IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2018

Project-Team MULTISPEECH

Speech Modeling for Facilitating Oral-Based Communication

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

RESEARCH CENTER
Nancy - Grand Est

THEME
Language, Speech and Audio

Table of contents

Project-Team MULTISPEECH

Creation of the Team: 2014 July 01, updated into Project-Team: 2015 July 01

Keywords:

Computer Science and Digital Science:

A3.4.6. - Neural networks
A3.4.8. - Deep learning
A5.1.7. - Multimodal interfaces
A5.7. - Audio modeling and processing
A5.7.1. - Sound
A5.7.2. - Music
A5.7.3. - Speech
A5.7.4. - Analysis
A5.7.5. - Synthesis
A5.8. - Natural language processing
A5.9.1. - Sampling, acquisition
A5.9.2. - Estimation, modeling
A5.9.3. - Reconstruction, enhancement
A5.9.5. - Sparsity-aware processing
A5.10.2. - Perception
A5.11.2. - Home/building control and interaction
A6.2.4. - Statistical methods
A6.3.1. - Inverse problems
A6.3.5. - Uncertainty Quantification
A9.2. - Machine learning
A9.3. - Signal analysis
A9.4. - Natural language processing

Other Research Topics and Application Domains:

B4.3.3. - Wind energy
B8.1.2. - Sensor networks for smart buildings
B8.4. - Security and personal assistance
B9.1.1. - E-learning, MOOC
B9.2.1. - Music, sound
B9.2.2. - Cinema, Television
B9.5.1. - Computer science
B9.5.2. - Mathematics
B9.5.6. - Data science
B9.6.8. - Linguistics
B9.6.10. - Digital humanities

1. Team, Visitors, External Collaborators

Research Scientists

Denis Jouvet [Team leader, Inria, Senior Researcher, on secondment from Corps des Mines, HDR]
Anne Bonneau [CNRS, Researcher]
Antoine Deleforge [Inria, Researcher, from Apr 2018]
Dominique Fohr [CNRS, Researcher]
Yves Laprie [CNRS, Senior Researcher, HDR]
Emmanuel Vincent [Inria, Senior Researcher, HDR]
Md Sahidullah [Inria, Starting Research Position, from Sep 2018]

Faculty Members

Vincent Colotte [Univ de Lorraine, Associate Professor]
Irène Illina [Univ de Lorraine, Associate Professor, HDR]
Odile Mella [Univ de Lorraine, Associate Professor]
Slim Ouni [Univ de Lorraine, Associate Professor, HDR]
Agnès Piquard-Kipffer [Univ de Lorraine, Associate Professor]
Romain Serizel [Univ de Lorraine, Associate Professor]

Post-Doctoral Fellows

Elodie Gauthier [Univ de Lorraine, from Mar 2018, granted by ANR & Région Grand-Est]
Md Sahidullah [Inria, until Aug 2018, partially granted by Region Grand-Est]

PhD Students

Théo Biasutto-Lervat [Univ de Lorraine, granted by ANR]
Guillaume Carbajal [Invoxia]
Sara Dahmani [Univ de Lorraine]
Ken Déguernel [Inria, until Feb 2018, granted by ANR & Région Lorraine]
Ioannis Douros [Univ de Lorraine]
Adrien Dufraux [Facebook AI Research, from Nov. 2018]
Baldwin Dumortier [Inria, until Aug 2018, granted by Venathec]
Raphaël Duroselle [Ministère des Armées, from Sep 2018]
Mathieu Fontaine [Inria, granted by ANR & Région Lorraine]
Nicolas Furnon [Univ de Lorraine, from Oct 2018, granted by ANR]
Amal Houidhek [Univ de Lorraine and École Nationale d'Ingénieurs de Tunis, Tunisia]
Ajinkya Kulkarni [Univ de Lorraine, from Oct 2018]
Manuel Pariente [Univ de Lorraine, from Oct 2018]
Laureline Perotin [Orange Labs]
Sunit Sivasankaran [Inria, granted by ANR]
Anastasiia Tsukanova [Univ de Lorraine]
Nicolas Turpault [Inria, partially granted by Région Grand-Est]

Technical staff

Yassine Boudi [Inria]
Valérian Girard [Inria, from Sep 2018]
Valérian Girard [Univ de Lorraine, until Aug 2018]
Thomas Girod [Univ de Lorraine, from Apr 2018]
Mathieu Hu [Inria]
Aditya Nugraha [Inria, from Mar 2018 until Apr 2018]
Anne-Laure Piat-Marchand [Univ de Lorraine, from Mar 2018]

Interns

Badr Abdullah [Univ de Lorraine, from Feb 2018 until Jun 2018]
Faustine Bille [Univ de Lorraine, from May 2018 until Jul 2018]
Gabriel Daubenfeld [Univ de Lorraine, from Apr 2018 until Jun 2018]
Aymerik Diebold [Univ de Lorraine, from Apr 2018 until Jun 2018]
Lea Dussere [Univ de Lorraine, from Jun 2018 until Jul 2018]
Camille Gresselin [Univ de Lorraine, from Jun 2018 until Sep 2018]

Lena Joyeux [Univ de Lorraine, from Sep 2018]
Ajinkya Kulkarni [Univ de Lorraine, from Feb 2018 until Jul 2018]
Constance Leroy [Univ de Lorraine, from Jun 2018 until Sep 2018]
Melody Louchard [Univ de Lorraine, from Jun 2018 until Jul 2018]
Quentin Millardet [Univ de Lorraine, from Apr 2018 until Jul 2018]
Noemie Nivert [Univ de Lorraine, from Jun 2018 until Jul 2018]
Claire Pinel [Univ de Lorraine, from Jun 2018 until Jul 2018]
Eva Sanudo [Univ de Lorraine, from Jun 2018 until Jul 2018]
Alexandre Sirjean [Univ de Lorraine, from May 2018 until Aug 2018]
Anais Valle [Univ de Lorraine, from Jun 2018 until Sep 2018]

Administrative Assistants
Hélène Cavallini [Inria]
Delphine Hubert [Univ de Lorraine]
Martine Kuhlmann [CNRS]

Visiting Scientists
Diego Di Carlo [Inria, Team PANAMA, Rennes, from Oct 2018 until Dec 2018]
Lou Lee [Univ de Lorraine, from Oct 2018]

External Collaborators
Nathan Libermann [Inria, Team PANAMA, Rennes]
Imene Zangar [École Nationale d'Ingénieurs de Tunis, Tunisia]

2. Overall Objectives

2.1. Overall Objectives

The goal of the project is the modeling of speech for facilitating oral-based communication. The name MULTISPEECH comes from the following aspects that are particularly considered:

- **Multisource aspects** - which means dealing with speech signals originating from several sources, such as speaker plus noise, or overlapping speech signals resulting from multiple speakers; sounds captured from several microphones are also considered.
- **Multilingual aspects** - which means dealing with speech in a multilingual context, as for example for computer assisted language learning, where the pronunciations of words in a foreign language (i.e., non-native speech) is strongly influenced by the mother tongue.
- **Multimodal aspects** - which means considering simultaneously the various modalities of speech signals, acoustic and visual, in particular for the expressive synthesis of audio-visual speech.

The project is organized along the three following scientific challenges:

- **The explicit modeling of speech.** Speech signals result from the movements of articulators. A good knowledge of their position with respect to sounds is essential to improve, on the one hand, articulatory speech synthesis, and on the other hand, the relevance of the diagnosis and of the associated feedback in computer assisted language learning. Production and perception processes are interrelated, so a better understanding of how humans perceive speech will lead to more relevant diagnoses in language learning as well as pointing out critical parameters for expressive speech synthesis. Also, as the expressivity translates into both visual and acoustic effects that must be considered simultaneously, the multimodal components of expressivity, which are both on the voice and on the face, are addressed to produce expressive multimodal speech.

- **The statistical modeling of speech.** Statistical approaches are common for processing speech and they achieve performance that makes possible their use in actual applications. However, speech recognition systems still have limited capabilities (for example, even if large, the vocabulary is limited) and their performance drops significantly when dealing with degraded speech, such as noisy or reverberated signals, distant microphone recording and spontaneous speech. Source separation and speech enhancement approaches are investigated as a way of making speech recognition systems more robust. Handling new proper names is an example of critical aspect that is tackled, along with the use of statistical models for speech-text automatic alignment and for speech production.
- **The estimation and the exploitation of uncertainty in speech processing.** Speech signals are highly variable and often disturbed with noise or other spurious signals (such as music or undesired extra speech). In addition, the output of speech enhancement and of source separation techniques is not exactly the accurate "clean" original signal, and estimation errors have to be taken into account in further processing. Hence, one goal consists to compute and handle the uncertainty of the reconstructed signal provided by source separation approaches. Finally, MULTISPEECH also aims to estimate the reliability of phonetic segment boundaries and prosodic parameters for which no such information is currently available.

Although being interdependent, each of these three scientific challenges constitutes a founding research direction for the MULTISPEECH project. Consequently, the research program is organized along three research directions, each one matching a scientific challenge. A large part of the research is conducted on French speech data; English and German languages are also considered in speech recognition experiments and language learning. Adaptation to other languages of the machine learning based approaches is possible, depending on the availability of corresponding speech corpora. Most of our research on signal processing, speech recognition, speech synthesis, etc., relies on deep learning approaches.

3. Research Program

3.1. Explicit Modeling of Speech Production and Perception

Speech signals are the consequence of the deformation of the vocal tract under the effect of the movements of the articulators (jaw, lips, tongue, ...) to modulate the excitation signal produced by the vocal cords or air turbulence. These deformations are visible on the face (lips, cheeks, jaw) through the coordination of different orofacial muscles and skin deformation induced by the latter. These deformations may also express different emotions. We should note that human speech expresses more than just phonetic content, to be able to communicate effectively. In this project, we address the different aspects related to speech production from the modeling of the vocal tract up to the production of expressive audiovisual speech. Phonetic contrasts used by the phonological system of any language result from constraints imposed by the nature of the human speech production apparatus. For a given language these contrasts are organized so as to guarantee that human listeners can identify (categorize) sounds robustly. The study of the categorization of sounds and prosody thus provides a complementary view on speech signals by focusing on the discrimination of sounds by humans, particularly in the context of language learning.

3.1.1. Articulatory modeling

Modeling speech production is a major issue in speech sciences. Acoustic simulation makes the link between articulatory and acoustic domains. Unfortunately this link cannot be fully exploited because there is almost always an acoustic mismatch between natural and synthetic speech generated with an articulatory model approximating the vocal tract. However, the respective effects of the geometric approximation, of the fact of neglecting some cavities in the simulation, of the imprecision of some physical constants and of the dimensionality of the acoustic simulation are still unknown. Hence, the first objective is to investigate the origin of the acoustic mismatch by designing more precise articulatory models, developing new methods to acquire tridimensional Magnetic Resonance Imaging (MRI) data of the entire vocal tract together with

denoised speech signals, and evaluating several approaches of acoustic simulation. Up to now, acoustic-to-articulatory inversion has been addressed as an instantaneous problem, articulatory gestures being recovered by concatenating local solutions. The second objective is thus to investigate how more elaborated strategies (a syllabus of primitive gestures, articulatory targets...) can be incorporated in the acoustic-to-articulatory inversion algorithms to take into account dynamic aspects.

3.1.2. Expressive acoustic-visual synthesis

Speech is considered as a bimodal communication means; the first modality is audio, provided by acoustic speech signals and the second one is visual, provided by the face of the speaker. In our approach, the Acoustic-Visual Text-To-Speech synthesis (AV-TTS) is performed simultaneously with respect to its acoustic and visible components. A first AV-TTS system has been developed resulting in a talking head; the system relied on 3D-visual data and on an extension of our acoustic-unit concatenation text-to-speech synthesis system (SOJA). An important goal is to provide an audiovisual synthesis that is intelligible, both acoustically and visually. Thus, we continue working on adding visible components of the head through a tongue model and a lip model. Another challenging research goal is to add expressivity in the AV-TTS. The expressivity comes through the acoustic signal (prosody aspects) and also through head and eyebrow movements. One objective is to add a prosodic component in the TTS engine in order to take into account some prosodic entities such as emphasis (to highlight some important key words).

3.1.3. Categorization of sounds and prosody for native and non-native speech

Discriminating speech sounds and prosodic patterns is the keystone of language learning whether in the mother tongue or in a second language. This issue is associated with the emergence of phonetic categories, i.e., classes of sounds related to phonemes and prosodic patterns. The study of categorization is concerned not only with acoustic modeling but also with speech perception and phonology. Foreign language learning raises the issue of categorizing phonemes of the second language given the phonetic categories of the mother tongue. Thus, studies on the emergence of new categories, whether in the mother tongue (for people with language deficiencies) or in a second language, must rely upon studies on native and non-native acoustic realizations of speech sounds and prosody, and on perceptual experiments. Concerning prosody, studies are focused on native and non-native realizations of modalities (e.g., question, affirmation, command, ...), as well as non-native realizations of lexical accents and focus (emphasis). For language learning, the aim is to provide automatic feedback to language learners with respect to acquisition of prosody as well as acquisition of a correct pronunciation of the sounds of the foreign language. Concerning the mother tongue we are interested in the monitoring of the process of sound categorization in the long term (mainly at primary school) and its relation with the learning of reading and writing skills, especially for children with language deficiencies.

3.2. Statistical Modeling of Speech

Whereas the first research direction deals with the physical aspects of speech and its explicit modeling, this second research direction investigates machine learning-based approaches for handling speech data. Acoustic models are used to represent the pronunciation of the sounds or other acoustic events such as noise. Whether they are used for source separation, for speech recognition, for speech transcription, or for speech synthesis, the achieved performance strongly depends on the accuracy of these models. At the linguistic level, MULTISPEECH investigates models for handling the context (thus going beyond the few preceding words of the n-gram models) and evolutive lexicons necessary when dealing with diachronic audio documents. With respect to the generation of speech signals, MULTISPEECH considers parametric speech synthesis applied to expressive multimodal speech synthesis.

3.2.1. Source separation

Acoustic modeling is a key issue for automatic speech recognition. Despite the progress made for many years, speech recognition performance depends on the quality of the input speech signals, and performance degrades quickly with noisy or reverberated signals. Accurate signal enhancement techniques are therefore essential to increase the robustness of both automatic speech recognition and speech-text alignment systems to

noise and non-speech events. In MULTISPEECH, focus is set on source separation techniques using multiple microphones and/or models of non-speech events. Some of the challenges include getting the most of the new modeling frameworks based on alpha-stable distributions and on deep neural networks, combining them with established spatial filtering approaches, modeling more complex properties of speech and audio sources (phase, inter-frame and inter-frequency properties), and exploiting large data sets of speech, noise, and acoustic impulse responses to automatically discover new models. Beyond the definition of such models, one difficulty is to design scalable estimation algorithms robust to overfitting, to integrate them into the recently developed FASST [6] and KAM software frameworks if relevant, and to develop new software frameworks otherwise.

3.2.2. Ambient sounds detection and classification

We are constantly surrounded by a complex audio stream carrying information about our environment. Hearing is a privileged way to detect and identify events that may require quick action (ambulance siren, baby cries...). Indeed, audition offers several advantages compared to vision: it allows for omnidirectional detection, up to a few tens of meters and independently of the lighting conditions. For these reasons, automatic audio analysis has become increasingly popular over the past few years. Yet, machines are still limited to detecting and classifying a few tens of sound event classes while human can generally recognize a few thousand. Besides, current algorithms rely heavily on the availability of annotated data that are extremely costly to obtain. In MULTISPEECH we focus on developing new methods, independent of applications, that would enable the detection of thousands of audio events from a little amount of annotated data while being robust to "out-of-the lab" conditions.

3.2.3. Linguistic modeling

MULTISPEECH investigates lexical and language models in speech recognition with a focus on improving the processing of proper names and of spontaneous speech. Proper names are relevant keys in information indexing, but are a real problem in transcribing many diachronic spoken documents which refer to data, especially proper names, that evolve over time. This leads to the challenge of dynamically adjusting lexicons and language models through the use of the context of the documents or of some relevant external information. We also investigate language models defined on a continuous space (through neural network based approaches) in order to achieve a better generalization on unseen data, and to model long-term dependencies. We also want to introduce into these models additional relevant information such as linguistic features, semantic relation, topic or user-dependent information. Other topics are spontaneous speech, for which utterances are often ill-formed and frequently contain disfluencies (hesitations, repetitions, ...) that degrade speech recognition performance, and pronunciation lexicons which are critical especially when dealing with non-native speech and foreign names.

3.2.4. Speaker identification

Speaker identification is the task that consists in identifying a person based on a voice recording. It has recently been deployed in several real-world application including secured access to bank services via telephone or internet. However, identification based solely on voice remains a modality with limited reliability under real conditions including several acoustic perturbations (noise, reverberation...) when the speaker might not be cooperative (a limited amount of data is available). In MULTISPEECH we focus on exploring new approaches exploiting multichannel speech enhancement techniques and uncertainty propagation to improve the performance of speaker identification systems in real conditions and with short speech utterances.

3.2.5. Speech generation by statistical methods

Over the last few years parametric speech synthesis has emerged as an alternative to corpus-based speech synthesis. The announced advantages of the parametric speech synthesis are the possibility to deal with small amounts of speech resources and the flexibility for adapting models (for new emotions or new speakers). MULTISPEECH investigates parametric approaches (currently based on deep learning) to produce expressive audio-visual speech. Also, in the context of acoustic feedback in foreign language learning, voice modification approaches are studied to modify the learner's (or teacher's) voice in order to emphasize the difference between the learner's acoustic realization and the expected realization.

3.3. Uncertainty Estimation and Exploitation in Speech Processing

This axis focuses on the uncertainty associated with some processing steps. Uncertainty stems from the high variability of speech signals and from imperfect models. For example, enhanced speech signals resulting from source separation are not exactly the clean original speech signals. Words or phonemes resulting from automatic speech recognition contain errors, and the phone boundaries resulting from an automatic speech-text alignment are not always correct, especially in acoustically degraded conditions. Hence the goal of estimating the reliability and/or the uncertainty on the results.

3.3.1. Uncertainty and acoustic modeling

Because small distortions in the separated source signals can translate into large distortions in the cepstral features used for speech recognition, this limits the recognition performance on noisy data. One way to address this issue is to estimate the uncertainty of the separated sources in the form of their posterior distribution and to propagate this distribution, instead of a point estimate, through the subsequent feature extraction and speech decoding stages. MULTISPEECH seeks to provide more accurate estimates of the posterior distribution of the separated source signals accounting for, e.g., posterior correlations over time and frequency which have not been considered so far. Mappings learned on training data and fusion of multiple uncertainty estimators are also explored. The estimated uncertainties are then exploited for acoustic modeling in speech recognition and, in the future, also for speech-text alignment. This approach may later be extended to the estimation of the resulting uncertainty of the acoustic model parameters and of the acoustic scores themselves.

3.3.2. Uncertainty and phonetic segmentation

The accuracy of the phonetic segmentation is important in several cases, as for example for the computation of prosodic features, for avoiding incorrect feedback to the learner in computer assisted foreign language learning, or for the post-synchronization of speech with face/lip images. Currently the phonetic boundaries obtained are quite correct on good quality speech, but the precision degrades significantly on noisy and non-native speech. Phonetic segmentation aspects need to be investigated, both in speech recognition (i.e., spoken text unknown) and in forced alignment (i.e., when the spoken text is known). In the same way that combining several speech recognition outputs leads to improved speech recognition performance, MULTISPEECH plans to investigate the combination of several speech-text alignments as a way of improving the quality of speech-text alignment and of getting information on the reliability of the phonetic boundaries. When segmenting speech corpora, knowing the reliability of the boundaries will help deciding which parts of the corpora need to be manually checked and corrected, thus avoiding an exhaustive checking of the whole corpus.

3.3.3. Uncertainty and prosody

Prosody information is also investigated as a means for structuring speech data (determining sentence boundaries, punctuation. . .) possibly in addition to syntactic dependencies. Structuring automatic transcription output is important for further exploitation of the transcription results such as easier reading after the addition of punctuation, or exploitation of full sentences in automatic translation. Prosody information is also necessary for determining the modality of the utterance (question or not), as well as determining accented words. Prosody information comes from the fundamental frequency, the duration of the sounds and their energy. Any error in estimating these parameters may lead to a wrong decision. MULTISPEECH aims at estimating the uncertainty of the duration of the phones (see uncertainty of phonetic boundaries above) and of the fundamental frequency estimates, as well as how this uncertainty shall be propagated in the detection of prosodic phenomena such as accented words, utterance modality, or determination of the structure of the utterance.

4. Application Domains

4.1. Introduction

Approaches and models developed in the MULTISPEECH project are intended to be used for facilitating oral communication in various situations through enhancements of the communication channels, either

directly via automatic speech recognition or speech production technologies, or indirectly, thanks to computer assisted language learning. Applications also include the usage of speech technologies for helping people in handicapped situations or for improving their autonomy. Foreseen application domains are related to computer assisted learning, health and autonomy (more precisely aided communication and monitoring), annotation and processing of spoken documents, and multimodal computer interaction.

4.2. Computer Assisted Learning

Although speaking seems quite natural, learning foreign languages, or learning the mother tongue for people with language deficiencies, represents critical cognitive stages. Hence, many scientific activities have been devoted to these issues either from a production or a perception point of view. The general guiding principle with respect to computer assisted mother or foreign language learning is to combine modalities or to augment speech to make learning easier. Based upon a comparison of the learner's production to a reference, automatic diagnoses of the learner's production can be considered, as well as perceptual feedback relying on an automatic transformation of the learner's voice. The diagnosis step strongly relies on the studies on categorization of sounds and prosody in the mother tongue and in the second language. Furthermore, making a reliable diagnosis on each individual utterance is still a challenge, which requires a temporally accurate segmentation of the speech utterance into phones; this explains why accurate segmentation of speech is an important topic in the field of acoustic speech modeling.

4.3. Aided Communication and Monitoring

A foreseen application aims at improving the autonomy of elderly or disabled people, and fits with smartroom applications. In a first step, source separation techniques should help for locating and monitoring people through the detection of sound events inside apartments. In a longer perspective, adapting speech recognition technologies to the voice of elderly people should also be useful for such applications, but this requires the recording of adequate databases. Sound monitoring in other application fields (security, environmental monitoring) can also be envisaged.

4.4. Annotation and Processing of Spoken Documents and Audio Archives

A first type of annotation consists in transcribing a spoken document in order to get the corresponding sequences of words, with possibly some complementary information, such as the structure (punctuation) or the modality (affirmation/question) of the utterances to make the reading and understanding easier. Typical applications of the automatic transcription of radio or TV shows, or of any other spoken document, include making possible their access by deaf people, as well as by text-based indexing tools.

A second type of annotation is related to speech-text alignment, which aims at determining the starting and ending times of the words, and possibly of the sounds (phonemes). This is of interest in several cases such as for annotating speech corpora for linguistic studies, and for synchronizing lip movements with speech sounds (for example, for avatar-based communications). Although good results are currently achieved on clean data, automatic speech-text alignment needs to be improved for properly processing noisy spontaneous speech data and needs to be extended to handle overlapping speech.

Large audio archives are important for some communities of users, e.g., linguists, ethnologists or researchers in digital humanities in general. In France, a notorious example is the "Archives du CNRS — Musée de l'homme", gathering about 50,000 recordings dating back to the early 1900s. When dealing with very old recordings, the practitioner is often faced with the problem of noise. This stems from the fact that a lot of interesting material from a scientific point of view is very old or has been recorded in very adverse noisy conditions, so that the resulting audio is poor. The work on source separation can lead to the design of semi-automatic denoising and enhancement features, that would allow these researchers to significantly enhance their investigation capabilities, even without expert knowledge in sound engineering.

Finally, there is also a need for speech signal processing techniques in the field of multimedia content creation and rendering. Relevant techniques include speech and music separation, speech equalization, prosody modification, and speaker conversion.

4.5. Multimodal Computer Interactions

Speech synthesis has tremendous applications in facilitating communication in a human-machine interaction context to make machines more accessible. For example, it started to be widely common to use acoustic speech synthesis in smartphones to make possible the uttering of all the information. This is valuable in particular in the case of handicap, as for blind people. Audiovisual speech synthesis, when used in an application such as a talking head, i.e., virtual 3D animated face synchronized with acoustic speech, is beneficial in particular for hard-of-hearing individuals. This requires an audiovisual synthesis that is intelligible, both acoustically and visually. A talking head could be an intermediate between two persons communicating remotely when their video information is not available, and can also be used in language learning applications as vocabulary tutoring or pronunciation training tool. Expressive acoustic synthesis is of interest for the reading of a story, such as audiobook, to facilitate the access to literature (for instance for blind people or for illiterate people).

5. Highlights of the Year

5.1. Highlights of the Year

E. Vincent has co-edited a 500-page book on audio source separation and speech enhancement, which provides a unifying view of various established and recent methods [64].

5.1.1. Awards

2018 ISCA Award for the best paper published in *Computer Speech and Language* (2013–2017) [1].

Best paper award of MISSI 2018 (11th International Conference on Multimedia and Network Information Systems) [44].

BEST PAPERS AWARDS:

[1]
J. BARKER, E. VINCENT, N. MA, H. CHRISTENSEN, P. GREEN. *The PASCAL CHiME Speech Separation and Recognition Challenge*, in "Computer Speech and Language", February 2013, vol. 27, n^{o} 3, pp. 621-633 [*DOI :* 10.1016/J.CSL.2012.10.004], https://hal.inria.fr/hal-00743529

[44]
K. SMAÏLI, D. FOHR, C. GONZÁLEZ-GALLARDO, M. GREGA, L. JANOWSKI, D. JOUVET, A. KOMOROWSKI, A. KOZBIAL, D. LANGLOIS, M. LESZCZUK, O. MELLA, M. A. MENACER, A. MENDEZ, E. LINHARES PONTES, E. SANJUAN, D. SWIST, J.-M. TORRES-MORENO, B. GARCIA-ZAPIRAIN. *A First Summarization System of a Video in a Target Language*, in "MISSI 2018 - 11th edition of the International Conference on Multimedia and Network Information Systems", Wrocław, Poland, September 2018, pp. 1-12, https://hal.archives-ouvertes.fr/hal-01819720

6. New Software and Platforms

6.1. dnnsep

Multichannel audio source separation with deep neural networks
KEYWORDS: Audio - Source Separation - Deep learning

SCIENTIFIC DESCRIPTION: dnnsep is the only source separation software relying on multichannel Wiener filtering based on deep learning. Deep neural networks are used to initialize and reestimate the power spectrum of the sources at every iteration of an expectation-maximization (EM) algorithm. This results in state-of-the-art separation quality for both speech and music.
FUNCTIONAL DESCRIPTION: Combines deep neural networks and multichannel signal processing for speech enhancement and separation of musical recordings.

RELEASE FUNCTIONAL DESCRIPTION: This version derives from version 1.0 (not 1.9). Differences concerns the use of a bidirectional long short-term memory (BLSTM) neural network, smoothing of the multichannel Wiener filter (MWF) over time and frequency, usage of the principal component of the MWF filter, adding a new generalized eigenvector beamformer with blind analytical normalization (GEVB) filter, and normalizing the training and test signals.

- Participants: Aditya Nugraha, Emmanuel Vincent and Antoine Liutkus
- Contact: Emmanuel Vincent

6.2. Dynalips-Player

High realistic lip synchronization for 3d animated characters
KEYWORDS: 3D animation - Graphics - Speech Synthesis
FUNCTIONAL DESCRIPTION: Dynalips provides a solution to synchronize precisely and automatically the movements of the lips of a 3D character with speech (we address 3D animation movies and video games). We have developed a demonstrator that illustrates the whole process: from audio + text to the generation of the animation trajectory, and controlling the animation of a 3D model (e.g. an avatar). The demonstrator is composed mainly by the player developed in Unity 3D (but can be used with any other system) and plays the animation synchronously with speech in realtime. It is possible to generate an animation for Autodesk Maya 3D.
NEWS OF THE YEAR: The whole lip-sync demonstrator is fully operational. From text and recorded speech, the system allows animating two different 3D models. The player is running on Unity 3D.

- Partners: Université de Lorraine - Sayens (SATT Grand Est)
- Contact: Slim Ouni
- URL: http://www.dynalips.com

6.3. KATS

Kaldi-based Automatic Transcription System
KEYWORD: Speech recognition
FUNCTIONAL DESCRIPTION: KATS is a multipass system for transcribing audio data, and in particular radio or TV shows in French, English or Arabic. It is based on the Kaldi speech recognition tools. It relies on Deep Neural Network (DNN) modeling for speech detection and acoustic modeling of the phones (speech sounds). Higher order statistical language models and recurrent neural network language models can be used for improving performance through rescoring of multiple hypotheses.
NEWS OF THE YEAR: New models have been trained for British English and evaluated on MGB data

- Contact: Dominique Fohr

6.4. VisArtico

Visualization of multimodal speech data
KEYWORDS: Data visualization - 3D movement - Speech processing - Videos

SCIENTIFIC DESCRIPTION: VisArtico is a multimodal data visualization software acquired by several systems: articulograph, motion capture, depth camera. This software makes it possible to visualize the positions of real or virtual sensors and to animate them simultaneously with acoustics. Regarding the articulatory data, the user has the possibility to visualise the contour of the tongue and the lips. It also makes it possible to find the midsagittal plane of the speaker, and to deduce the position of the palate, if this information is absent during the acquisition. The software makes it possible to display the segmentation at the level of sentences, words or phonemes. The goal is to provide an effective multimodal data visualization tool that can be useful to anyone studying speech production, audio-visual synthesis, or animation in a more general way.
FUNCTIONAL DESCRIPTION: VisArtico is a user-friendly software which allows visualizing multimodal data acquired by several systems : an articulograph (AG500, AG501 or NDI Wave), motion capture system, depth camera. This visualization software has been designed so that it can directly use the data provided by the different systems to display the spatial and temporal positions of the sensors (real and virtual), synchronized with the corresponding acoustic recordings. Moreover, for articulatory data, VisArtico not only allows viewing the sensors but also enriches the visual information by indicating clearly and graphically the data for the tongue, lips and jaw. Finally, it is possible to generate a movie for any articulatory-acoustic sequence. This software can be useful for researchers in speech production, audiovisual speech synthesis or articulatory speech analysis.

RELEASE FUNCTIONAL DESCRIPTION: The main improvement in this version is the ability to view a video that was recorded along with the articulatory or motion capture data. The software also allows for automatic speech segmentation.
NEWS OF THE YEAR: This year, we have added the possibility to visualize a video simultaneously with the multimodal data and the acoustic data. Several bugs have been fixed.

- Participants: Ilef Ben Farhat, Loïc Mangeonjean and Slim Ouni
- Partners: CNRS - Université de Lorraine
- Contact: Slim Ouni
- Publication: VisArtico: a visualization tool for articulatory data
- URL: http://visartico.loria.fr

6.5. Xarticulators

KEYWORDS: Medical imaging - Natural language processing
FUNCTIONAL DESCRIPTION: The Xarticulators software is intended to delineate contours of speech articulators in X-ray and MR images, construct articulatory models and synthesize speech from X-ray films. This software provides tools to track contours automatically, semi-automatically or by hand, to make the visibility of contours easier, to add anatomical landmarks to speech articulators and to synchronize images with the sound. In addition we also added the possibility of processing digitized manual delineation results made on sheets of papers when no software is available. Xarticulators also enables the construction of adaptable linear articulatory models from the X-ray or MR images and incorporates acoustic simulation tools to synthesize speech signals from the vocal tract shape. Recent work was on the possibility of synthesizing speech from 2D-MRI films, and on the construction of better articulatory models for the velum, lips and epiglottis.

RELEASE FUNCTIONAL DESCRIPTION: The new version allows MRI films to be processed and, above all, it offers a better transition from the shape of the vocal tract to the area function, which corresponds to an approximation of the vocal tract using a series of elementary tubes from the glottis to the lips.
NEWS OF THE YEAR: Improvements to the articulatory model which now takes into account "small" filiform articulators such as the epiglottis and velum using models using the central line. Improvement of the transition from the medio-sagittal 2D form of the vocal tract to the function of area.

- Contact: Yves Laprie
- Publication: Articulatory model of the epiglottis

7. New Results

7.1. Explicit Modeling of Speech Production and Perception

Participants: Anne Bonneau, Vincent Colotte, Denis Jouvet, Yves Laprie, Slim Ouni, Agnès Piquard-Kipffer, Théo Biasutto-Lervat, Sara Dahmani, Ioannis Douros, Valérian Girard, Thomas Girod, Anastasiia Tsukanova.

7.1.1. Articulatory modeling

7.1.1.1. Articulatory models and synthesis

Since articulatory modeling, i.e. representing the geometry of the vocal tract with a small number of parameters, is a key issue in articulatory synthesis the improvement of the articulatory models remains an important objective. This year we put emphasis on thin articulators as the epiglottis and velum. Indeed, the delineation of those contours often leads to erroneous transverse dimensions (too thin or too thick contours) which generates some artificial swelling deformations. Before the determination of the deformation modes, the central lines of the velum and epiglottis are extracted in the images use to build the model. The deformation modes thus only concern the central line, which prevents artificial swelling factors to emerge from the factor analysis. A reconstruction algorithm has been developed to obtain the contour from the central line.

7.1.1.2. Acoustic simulations

One of the issues in articulatory synthesis is to assess the impact of the geometric simplifications that are made on the vocal tract so as to enable faster acoustic simulation and to decrease the number of parameters required to approximate the vocal tract shape. The other issue concerns the impact of the plane wave assumption. The idea consists of comparing the signal or spectrum synthesized via numerical acoustic simulation against the one measured on a real human subject. However, this requires that both geometric and corresponding acoustic data are available at the same time. This can be achieved with MRI data when the acquisition duration is sufficiently short to allow the speaker to phonate the sound during the whole acquisition. The MRI acquisition protocol has thus been optimized on the new Siemens Prisma MRI machine of Nancy hospital so as to reduce the acquisition time to 7 seconds, which makes it possible for the subject to produce a sound throughout the acquisition. The acoustic simulation was achieved by using the Matlab K-wave package, either from the entire 3D volume extracted from the MRI data, or from the 2D shape extracted from the mid-sagittal plane. Several simplifications have been carried out (with or without the epiglottis, with or without the velum...) so as to assess their acoustic impacts. These simulations only concern vowels because these sounds can be sustained by subjects and the MRI machine noise does not change the position of formant frequencies dramatically. This work has been carried out in cooperation with IADI laboratory.

7.1.1.3. Exploitation of dynamic articulatory data

The size of the dynamic database (recorded last year in the Max Planck Institute for Biophysical Chemistry in Göttingen), in the form of MRI films of the mid-sagittal plane acquired at 55 Hz, is about 200.000 images. Even if the long term objective is to exploit the whole database, efforts were dedicated to manual delineation of contours in some films with the idea of using those data to train a machine learning technique. Several students were trained, and in total more than 1000 images have been delineated. The corresponding films have been exploited to achieve articulatory copy synthesis by improved acoustic simulations developed last year.

7.1.1.4. Acoustic-to-articulatory inversion

Deriving articulatory dynamics from the acoustic speech signal is a recurrent topic in our team. This year, we have investigated whether it is possible to predict articulatory dynamics from phonetic information without having the acoustic speech signal. The input data may be considered as not sufficiently rich acoustically, as there is probably no explicit coarticulation information, but we expect that the phonetic sequence provides compact yet rich knowledge. We have experimented a recurrent neural network architecture, where we have trained the model with an electromagnetic articulography (EMA) corpus, and have obtained good performances similar to the state-of-the-art articulatory inversion from line spectral frequencies (LSF) features [21].

7.1.2. Expressive acoustic and visual synthesis

7.1.2.1. Expressive speech

A comparison between emotional speech and neutral speech has been carried on using a small corpus of acted speech. The analysis was focused on the way pronunciations and prosodic parameters are modified in emotional speech, compared to neutral style [20].

Experiments with deep learning-based approaches for expressive speech synthesis are described in 7.2.4.2.

7.1.2.2. Expressive audiovisual synthesis and lipsync

This year, we have acquired audiovisual 3D corpus (using the optitrack system, using 8 cameras) for a set of emotions acted by a professional actress. We recorded 6 basic emotions: joy, fear, disgust, sadness, anger, surprise; in addition to neutral speech. The corpus contains 5000 utterances (2000 utterances for the neutral speech and 500 utterances per emotion). The visual and acoustic data have been processed, segmented and labeled spatially and temporally. An important aspect of the work was to study the evaluation of the quality of the animation of a 3D talking head where the animation is generated from the acquired 3D data. For this purpose, we studied the relevance of root mean square error (RMSE) measure which is classically used to evaluate the error of the prediction. Our preliminary results confirmed that RMSE can be irrelevant in our field, as we may not reach critical articulatory target, and we still obtain very low RMSE. Thus the audiovisual intelligibility of the system would be low. To improve the results, we have worked on improving the 3D model controls using better key-shapes and reduced redundant and confusing blendshapes.

The processed neutral-speech data have been used to train a deep neural network to predict from speech and linguistic information the trajectories of the animation controls of the talking head, which is the core of the lipsync system. We have also used this expressive-speech data to train a DNN-based TTS to synthesize expressive audiovisual speech from text. Currently, we are performing extensive testing and validation of the results.

7.1.3. Categorization of sounds and prosody for native and non-native speech

7.1.3.1. Visual clues in speech perception and production

We continue our research focused on the importance of multimodal speech combining oral and visual clues. We investigated identification and production of morpho-syntactic skills in ten deaf children (severe with cochlear implant using French cued-speech LPC - *Langue française Parlée Complétée*) and ten age-matched children with typical development. Our goal was to examine the production of morpho-syntactic structures in auditory channel versus audiovisual speech. Five conditions were observed: audiovisual conditions with a 3D avatar speaking or coding oral language with LPC versus a human speaker with or without LPC and auditory channel. We used the 3D avatar coding set up in the ADT Handicom project. Statistical analysis and interpretation of results is ongoing.

7.1.3.2. Reading and related skills norms

We set-up standardized norms on the development of reading and related skills in French: EVALEC Primaire software (in collaboration with the LPC - *Laboratoire de Psychologie Cognitive*, UMR 7290, *Aix-Marseille Université*). This year, LPC collected new data at the end of grade 5 (about 100 children) and added them to those previously collected at the end of grades 1–4, about 100 children for each level [69]. EVALEC primaire software includes five tests focused on written word processing, recording both accuracy scores and processing time (time latency and vocal response duration for the reading aloud tests). EVALEC primaire software also includes tests of phonemic and syllabic awareness, phonological short-term memory, and rapid naming. These data would allow researchers and speech therapists to assess the reading and reading-related skills of dyslexic children as compared to average readers.

7.1.3.3. Analysis of non-native pronunciations

We have examined the effects of L1/L2 interferences at the segmental level, and of the lack of fluency at the sentence level, on the realizations of French final fricatives by German learners. Due to L1/L2 interference, German speakers tend to devoice French final fricatives. A well-known effect of the lack of L2 mastering is the decrease of the speech articulation rate, which lengthens the average duration of segments. In order to better apprehend the impact of categorization and fluency, we selected four series of consonants from the IFCASL corpus, i.e. voiced and unvoiced fricatives uttered by French native and German non-native speakers. The realizations of French unvoiced consonants uttered by German speakers are essentially dependent on fluency, whereas the realizations of voiced consonants by the same speakers are dependent on both fluency and categorization. We evaluated a set of acoustic cues related to the voicing distinction -including consonant duration and periodicity-, and submitted the data to a hierarchical clustering analysis. Results, discussed as a function of speaker's level and prosodic boundaries, confirmed the mutual importance of fluency and segmental categorization on non-native realizations [22].

Within the METAL project, work is on-going for integrating speech processing technology in an application to help learning foreign language and for experimenting it with middle and high school students learning German. This includes tutoring aspects using a talking head to show proper articulation of words and sentences; as well as using automatic tools derived from speech recognition technology, for analyzing student pronunciations. Preliminary experiments have shown the poor quality of speech signals recorded from groups of students in classrooms.

7.2. Statistical Modeling of Speech

Participants: Vincent Colotte, Antoine Deleforge, Dominique Fohr, Irène Illina, Denis Jouvet, Odile Mella, Romain Serizel, Emmanuel Vincent, Md Sahidullah, Guillaume Carbajal, Ken Déguernel, Diego Di Carlo, Adrien Dufraux, Raphaël Duroselle, Mathieu Fontaine, Nicolas Furnon, Amal Houidhek, Ajinkya Kulkarni, Nathan Libermann, Aditya Nugraha, Manuel Pariente, Laureline Perotin, Sunit Sivasankaran, Nicolas Turpault, Imene Zangar.

7.2.1. Source localization and separation

Emmanuel Vincent has co-edited a 500-page book on audio source separation and speech enhancement, which provides a unifying view of array processing, matrix factorization, deep learning and other methods, with application to speech and music [64]. We also contributed to five chapters in that book [60], [62], [59], [54], [61] and three chapters in another book [53], [56], [55].

7.2.1.1. Source localization

In multichannel scenarios, source localization and source separation are tightly related tasks. We introduced the real and imaginary parts of the acoustic intensity vector in each time-frequency bin as suitable input features for deep learning based speaker localization [37]. We analyzed the inner working of the neural network using a methodology called layerwise relevance propagation, which points the time-frequency bins on which the network relies to output a given location [68]. We defined a new task called text-informed speaker localization, which consists of localizing the speaker uttering a known word or sentence such as the wake-up word of a hands-free voice command system in a situation when other speakers are overlapping. We proposed a method to address this task, where a phonetic alignment is obtained, converted into an estimated time-frequency mask, and fed to a convolutional neural network together with interchannel phase difference features in order to localize the desired speaker [43]. We published a new dataset using a microphone array embedded in an unmanned aerial vehicle in [45], organized an international sound source localization challenge associated to this dataset and participated to the 2018 LOCATA sound source localization challenge. We published a book chapter on audio-motor integration, showing an application to sound source localization with robots [52].

7.2.1.2. Room acoustics modeling

In a given room, each possible position of the microphones and the sources corresponds to different room transfer functions. The goal of room acoustic modeling is to model the manifold formed by these transfer functions. Past studies have focused on learning a supervised mapping between the relative transfer function and the source location for localization purposes. We introduced the reverse task consisting of learning a mapping between the source location and the corresponding relative transfer function, which may be used as a prior on the relative transfer function for source separation purposes. We proposed a semi-supervised algorithm to learn this mapping in a situation when the location of each relative transfer function measurement is not precisely known [48]. We also started investigating the estimation and modeling of early acoustic echoes. In [39] we showed how their knowledge could improve performance of sound source separation algorithms. In [36] we proposed a new method to estimate them blindly from multichannel recordings with much higher precision than conventional blind channel identification methods.

7.2.1.3. Deep neural models for source separation and echo suppression

We pursued our research on the use of deep learning for multichannel source separation [5]. We introduced a method that exploits knowledge of the source locations in order to estimate multichannel Wiener filters for two or more sources [38]. We explored several variants of the multichannel Wiener filter, which turned out to result in better speech recognition performance on the CHiME-3 dataset [17]. We also used deep neural networks for reducing the residual nonlinear echo after linear acoustic echo cancellation [23] and started extending this approach to joint reverberation, echo, and noise reduction. Finally, we recently started exploring the case where the microphones composing a multichannel array are not distributed according to a predefined geometry and do not have a common sampling clock.

7.2.1.4. Alpha-stable modeling of audio signals

This year, our work on heavy tails distribution has witnessed a significant advance with the development of a multichannel model that is able to account for the inter-channel delays and time difference of arrivals in an alpha-stable framework, hence benefiting from the inherent robustness of such distributions. This work has been submitted to the IEEE transactions on Signal Processing by Mathieu Fontaine and is still under review. Its main applications are: i/ the separation of multichannel sources, for which we have demonstrated a superiority with respect to the multichannel Wiener filter in the oracle setting, and ii/ localizations of heavy tailed sources, where we worked on the theoretical foundations

7.2.1.5. Beyond Gaussian modeling of audio signals

The team has investigated a number of alternative probabilistic models to the symmetric local complex Gaussian (LCG) model for audio source separation. An important limit of LCG is that most signals of interest such as speech or music do not exhibit Gaussian distributions but heavier-tailed ones due to their important dynamic. In [31] we proposed a new sound source separation algorithm using heavy-tailed alpha stable priors for source signals. Experiments showed that it outperformed baseline Gaussian-based methods on under-determined speech or music mixtures. Another limitation of LCG is that it implies a zero-mean complex prior on source signals. This induces a bias towards low signal energies, in particular in under-determined settings. With the development of accurate magnitude spectrogram models for audio signals using deep neural networks, it becomes desirable to use probabilistic models enforcing stronger magnitude priors and better accounting for phases. In [35], we presented the BEADS (Bayesian Expansion Approximating the Donut Shape) model. The prior considered is a mixture of isotropic Gaussians regularly placed on a zero-centered complex circle. We showed it outperformed LCG on an informed source separation task.

7.2.1.6. Interference reduction

Our work on interference reduction focused this year in scaling our previous work to full-length recording. This has been achieved thanks to a new method we proposed, which estimates the interference reduction parameters based on random projections of the full length recordings [25]. This technique scales linearly with the duration of the recording, making it usable in real-world use-cases.

The book chapter we published on audio-motor integration, shows an application to ego-noise reduction for robots [52]. In the context of robotics, ego-noise refers to the acoustic noise produced in a robot's microphones by its own movement.

7.2.2. Acoustic modeling

7.2.2.1. Robust acoustic modeling

Achieving robust speech recognition in reverberant, noisy, multi-source conditions requires both speech enhancement and separation and robust acoustic modeling. In order to motivate further work by the community, we created the series of CHiME Speech Separation and Recognition Challenges in 2011 [1]. We oversaw the collection of a new dataset sponsored by Google, which considers a 'dinner party' scenario. Twenty parties of four people, who know each other well, were recorded in their own homes using 2 binaural in-ear microphones per participant and 6 distant Kinects, for a total duration of about 50 h. We organized the CHiME-5 Challenge based on these data [19]. We also participated in the collection of two French datasets for ambient assisted living applications as part of the voiceHome [11] and VOCADOM [51] projects.

7.2.2.2. Ambient sounds

We are constantly surrounded by sounds and we rely heavily on these sounds to obtain important information about what is happening around us. Our team has been involved in the community on ambient sound recognition for the past few years. In collaboration with Johannes Kepler University (Austria) and Carnegie Mellon University (USA), we co-organized a task on large-scale sound event detection as part of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Challenge [40]. It focused on the problem of learning from audio segments that are either weakly labeled or not labeled, targeting domestic applications. In this context, we work on semi-supervised sampling strategies to create triplets (a triplet is composed of the current sample, a so-called positive sample from the same class as the current sample and a negative sample from a different class) and studied their application to train triplet networks for audio tagging.

7.2.2.3. Speech/Non-speech detection

Automatic Speech Recognition (ASR) of multimedia content such as videos or multi-genre broadcasting requires a correct extraction of speech segments. We explored the efficiency of deep neural models for speech/non-speech segmentation. We used a bidirectional LSTM model to obtain speech/non-speech probabilities and a decision module (4-state automaton with safety margins). Compared to a Gaussian Mixture Model (GMM) based speech/non-speech segmenter, the results achieved on the MGB British Challenge data, show a reduction of the ASR word error rate (23.7% versus 29.4%). We have also trained models for the Arabic and French languages.

7.2.2.4. Transcription systems

Within the AMIS project, speech recognition systems have been developed for the transcription of videos in French, English and Arabic. They have been integrated with other components (such as translation and summarization) to allow for the summarization of videos in a target language [44], [29], [28].

7.2.2.5. Speaker recognition

Speaker recognition is the task of recognizing a person from its voice. The performances of speaker recognition systems severely degrade due to several practical challenges such as the limited amount of speech data, real-world noises and spoofing. We explored the efficiency of DNN-based distance metric learning methods for speaker recognition in short duration conditions. Currently, we are developing a neural network architecture that gives phone-invariant speaker embeddings for robust speaker recognition. We also participated in the NIST speaker recognition evaluation 2018 as a part of the I4U consortium. The speaker recognition technology is vulnerable to spoofing attacks where mimicked voice, synthetic speech, or playback voice is used to get illegitimate access. We are investigating whether technology-assisted speaker selection can help in improving mimicry attack [67]. In [24], we proposed an enhanced baseline system for replay spoofing detection with ASVspoof 2017 dataset. In [26], we demonstrated that playback speech enhanced with DNN-based speech enhancement method can severely degrade the speaker recognition and countermeasure performance as compared to the conventional replay attacks with voice samples from covert recording. We also proposed

a common feature and back-end fusion scheme for the integration of spoofing countermeasures and speaker recognition [47]. Currently, we are co-organizing the third edition of automatic speaker verification spoofing challenge (ASVspoof 2019) where our newly developed cost function [32] will be adopted for the performance assessment of integrated systems. In the context of multimodal authentication with the voice as a modality, we investigated the optimization of speech features for audio-visual synchrony detection [41].

7.2.2.6. Language identification

With respect to language identification, the current research activity focuses on lightly supervised or unsupervised domain adaptation. The goal is to adapt a language identification system optimized for a given transmission channel to a new transmission channel.

7.2.3. Language modeling

7.2.3.1. Out-of-vocabulary proper name retrieval

Despite recent progress in developing Large Vocabulary Continuous Speech Recognition Systems (LVCSR), these systems suffer from Out-Of-Vocabulary words (OOV). In many cases, the OOV words are Proper Nouns (PNs). The correct recognition of PNs is essential for broadcast news, audio indexing, etc. We addressed the problem of OOV PN retrieval in the context of broadcast news LVCSR. We focused on dynamic (document dependent) extension of LVCSR lexicon. To retrieve relevant OOV PNs, we proposed to use a very large multipurpose text corpus: Wikipedia. This corpus contains a huge number of PNs. These PNs are grouped in semantically similar classes using word embedding. We used a two-step approach: first, we selected OOV PN pertinent classes with a multi-class Deep Neural Network (DNN). Secondly, we ranked the OOVs of the selected classes. The experiments on French broadcast news show that a bi-directional Gated Recurrent Unit model outperforms other studied models. Speech recognition experiments demonstrate the effectiveness of the proposed methodology [18].

7.2.3.2. Updating speech recognition vocabularies

Within the AMIS project, the update of speech recognition vocabularies has been investigated using web data collected over a time period similar to that of the collected videos, for three languages: French, English and Arabic. Results have been analyzed globally, and also with respect to names only. This analysis has shown the poor coverage of the names by the baseline lexicons, and has also demonstrated the benefits of the updated lexicons, both in term of WER reduction and OOV rate reduction [14].

7.2.3.3. Music language modeling

Similarly to speech, music involves several levels of information, from the acoustic signal up to cognitive quantities such as composer style or key, through mid-level quantities such as a musical score or a sequence of chords. The dependencies between mid-level and lower- or higher-level information can be represented through acoustic models and language models, respectively. Ken Déguernel defended his PhD on automatic music improvisation [10] and he proposed a polyphonic music improvisation approach that takes the structure of the musical piece at multiple time scales into account [12]. We also explored the ability of a conventional recurrent neural network with moving history to account for long-term dependencies in music melodies, and compared it with two new architectures with growing or parallel history [50].

7.2.3.4. Automatic detection of hate speech

Nowadays, Twitter, LinkedIn, Facebook and YouTube are very popular for communicating ideas, beliefs, feelings or any other form of information. At the same time, the dark side of these new technologies has led to an increase in hate speech or racism. Our work seeks to study hate speech in user-generated contents in France, which thus requires French resources. We plan to design a hate speech corpus and a lexicon in French; whereas such hate speech lexicons exist for other languages, no such tool can be found in French. We began, on English data, to develop a new methodology to automatically detect hate speech, based on machine learning and Neural Networks. Human detection of this material is unfeasible since the contents to be analyzed are huge. Current machine learning methods use only certain task specific features to model hate speech. We propose to develop an innovative approach to combine these pieces of information into a multi-feature approach so that the weaknesses of the individual features are compensated by the strengths of other features. We began a collaboration with the CREM laboratory in Metz and Saarland University.

7.2.4. Speech generation

7.2.4.1. Arabic speech synthesis

Work on Arabic speech synthesis was carried out within a CMCU PHC project with ENIT (École Nationale d'Ingénieurs de Tunis, Tunisia), using HMM and NN based approaches applied to Modern Standard Arabic language. Speech synthesis systems rely on a description of speech segments corresponding to phonemes, with a large set of features that represent phonetic, phonologic, linguistic and contextual aspects. When applied to Modern Standard Arabic, two specific phenomena have to be taken in account: vowel quantity and gemination. This year, we studied thoroughly the modeling of these phenomena. Results of objective and subjective evaluations showed that the use of a deep neural architecture in speech synthesis (more specifically in predicting the speech parameters) enhanced the accuracy of acoustic modelling so that the quality of generated speech is better than that of HMM-based speech synthesis [30], [13].

Deep neural network (DNN) approaches have been further investigated for the modeling of phoneme duration. According to the specific phenomena of the Arabic language, we proposed a class-specific modeling of the phoneme durations. An objective evaluation showed that the proposed approach leads to a more accurate modeling of the phoneme duration (compared to HMM-based or MERLIN DNN-based approaches) [49].

7.2.4.2. Expressive acoustic synthesis

Expressive speech synthesis using parametric approaches is constrained by the style of the speech corpus used. We carried out a preliminary study on developing expressive speech synthesis for a new speaker voice without requiring a specific recording of expressive speech by this new speaker. For that, we focused on deep neural network based layer adaptation for investigating the transfer the expressive characteristics to a new speaker for which only neutral speech data is available. Such transfer learning mechanism should accelerate the efforts towards exploiting existing expressive speech corpora. However, there is a trade-off between the knowledge transfer of expressivity characteristics and the retaining of the speaker's identity in the synthesized speech.

7.3. Uncertainty Estimation and Exploitation in Speech Processing

Participants: Irène Illina, Denis Jouvet, Emmanuel Vincent, Yassine Boudi, Baldwin Dumortier, Elodie Gauthier, Mathieu Hu, Lou Lee, Anne-Laure Piat-Marchand.

7.3.1. Uncertainty and acoustic modeling

7.3.1.1. Uncertainty in noise-robust speech and speaker recognition

In many real-world conditions, the target speech signal overlaps with noise and some distortion remains after speech enhancement. The framework of uncertainty decoding assumes that this distortion has a Gaussian distribution and seeks to estimate its covariance matrix and propagate it through the acoustic model for robust ASR [4]. We introduced new Gaussian mixture model-derived (GMMD) uncertainty features for robust DNN-based acoustic model training and decoding, which are computed as the difference between the closed-form GMM log-likelihoods obtained with vs. without uncertainty. We concatenated the GMMD features with conventional acoustic features and showed that they improve ASR performance on both the CHiME-2 and CHiME-3 datasets [15].

7.3.1.2. Uncertainty in other applications

Besides the above application, we finalized our exploration of uncertainty modeling for wind turbine control. Baldwin Dumortier defended his PhD thesis on this topic [9].

7.3.2. Uncertainty and phonetic segmentation

In the METAL project, experiments are planned to investigate further the use of speech technologies for foreign language learning in middle and high schools. Besides adapting acoustic models to teenager voices, current work investigates the reliability of speech technologies for analyzing student pronunciations, and for detecting miss-pronunciations. Also, besides making the pronunciation diagnostics more reliable, the aim is to elaborate robust strategies that will make it possible to handle sets of unreliable individual results, and still be able to provide a relevant feedback on recurrent miss-pronunciations.

7.3.3. Uncertainty and prosody

The analysis of prosodic correlates of discourse particles has continued. Some additional data has been annotated. The automatic word and phonetic segmentation of the discourse particles has been manually checked and corrected when necessary. Once more, this has shown that automatic segmentation is not perfect, especially on spontaneous speech recording in real conditions. For each discourse particle, prosodic characteristics of occurrences of each pragmatic function (conclusive, introductive, etc.) were automatically extracted. For each discourse particle and each pragmatic function, the most frequent F0 patterns were retained as the representative forms. Results show that a pragmatic function, common to several discourse particles, gives rise to a uniform prosodic marking [34].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Dolby

Company: Dolby (Spain)

Duration: Sep – Dec 2018

Participants: Antoine Liutkus (Inria Zenith), Emmanuel Vincent

Abstract: This contract aims to evaluate the feasibility of state-of-the-art source separation technology and related technologies for four use cases, and to identify those which could be commercially exploited, possibly after a follow-up R&D phase.

8.1.2. Honda Research Intitute Japan (first contract)

Company: Honda Research Intitute Japan (Japan)

Duration: Feb – Mar 2018

Participants: Aditya Nugraha, Romain Serizel, Emmanuel Vincent

Abstract: This contract targets collaborative research on multichannel speech and audio processing and eventual software licensing in order to enable voice-based communication in challenging noisy and reverberant conditions in which current hands-free voice-based interfaces perform poorly.

8.1.3. Honda Research Intitute Japan (second contract)

Company: Honda Research Intitute Japan (Japan)

Duration: Aug 2018 – Mar 2019

Participants: Nancy Bertin (CNRS - IRISA), Antoine Deleforge, Diego Di Carlo

Abstract: This is a follow-up contract, which also targets collaborative research on multichannel speech and audio processing and eventual software licensing in order to enable voice-based communication in challenging noisy and reverberant conditions in which current hands-free voice-based interfaces perform poorly.

8.1.4. Studio Maia

Company: Studio Maia SARL (France)

Other partners: Imaging Factory

Duration: Jul 2017 – March 2019

Participants: Yassine Boudi, Vincent Colotte, Mathieu Hu, Emmanuel Vincent

Abstract: This Inria Innovation Lab aims to develop a software suite for voice processing in the multimedia creation chain. The software is aimed at sound engineers and it will rely on the team's expertise in speech enhancement, robust speech and speaker recognition, and speech synthesis.

8.2. Bilateral Grants with Industry

8.2.1. Orange

Company: Orange SA (France)

Duration: Nov 2016 – Nov 2019

Participants: Laureline Perotin, Romain Serizel, Emmanuel Vincent

Abstract: This CIFRE contract funds the PhD thesis of Laureline Perotin with Orange Labs. Our goal is to develop deep learning based speaker localization and speech enhancement algorithms for robust hands-free voice command. We are especially targeting difficult scenarios involving several simultaneous speakers.

8.2.2. Invoxia

Company: Invoxia SAS (France)

Duration: Mar 2017 – Mar 2020

Participants: Guillaume Carbajal, Romain Serizel, Emmanuel Vincent

Abstract: This CIFRE contract funds the PhD thesis of Guillaume Carbajal. Our goal is to design a unified end-to-end deep learning based speech enhancement system that integrates all steps in the current speech enhancement chain (acoustic echo cancellation and suppression, dereverberation, and denoising) for improved hands-free voice communication.

8.2.3. Ministère des Armées

Company: Ministère des Armées (France)

Duration: Sep 2018 – Aug 2021

Participants: Raphaël Duroselle, Denis Jouvet, Irina Illina

Abstract: This contract corresponds to the PhD thesis of Raphaël Duroselle on the application of deep learning techniques for domain adaptation in speech processing.

8.2.4. Facebook

Company: Facebook AI Research (France)

Duration: Nov 2018 – Nov 2021

Participants: Adrien Dufraux, Emmanuel Vincent

Abstract: This CIFRE contract funds the PhD thesis of Adrien Dufraux. Our goal is to explore cost-effective weakly supervised learning approaches, as an alternative to fully supervised or fully unsupervised learning for automatic speech recognition.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. CPER LCHN

Project acronym: CPER LCHN

Project title: CPER “Langues, Connaissances et Humanités Numériques”

Duration: 2015-2020

Coordinator: Bruno Guillaume (LORIA) & Alain Polguère (ATILF)

Participants: Dominique Fohr, Denis Jouvet, Odile Mella, Yves Laprie

Abstract: The main goal of the project is related to experimental platforms for supporting research activities in the domain of languages, knowledge and numeric humanities engineering.

MULTISPEECH contributes to automatic speech recognition, speech-text alignment and prosody aspects.

9.1.2. CPER IT2MP

Project acronym: CPER IT2MP

Project title: CPER "Innovation Technologique Modélisation et Médecine Personalisée"

Duration: 2015-2020

Coordinator: Faiez Zannad (Inserm-CHU-UL)

Participants: Romain Serizel, Emmanuel Vincent

Abstract: The goal of the project is to develop innovative technologies for health, and tools and strategies for personalized medicine.

MULTISPEECH will investigate acoustic monitoring using an array of microphones.

9.1.3. Dynalips

Project title: Control of the movements of the lips in the context of facial animation for an intelligible lipsync.

Duration: February 2017 - August 2018

Coordinator: Slim Ouni

Participants: Valerian Girard, Slim Ouni

Funding: SATT

Abstract: We proposed in this project the development of tools of lipsync which, from recorded speech, provide realistic mechanisms of animating the lips. These tools are meant to be integrated into existing 3D animation software and existing game engines. One objective was that these lipsync tools fit easily into the production pipeline in the field of 3D animation and video games. The goal of this maturation was to propose a product ready to be exploited in the industry whether by the creation of a start-up or by the distribution of licenses.

A first prototype of the lipsync system has been developed for French. From audio and text, the system allows animating a 3D model of the face (an avatar) realistically. This work has been presented at Annecy International Animation Film Festival.

9.2. National Initiatives

9.2.1. ANR DYCI2

Project acronym: DYCI2 (http://repmus.ircam.fr/dyci2/)

Project title: Creative Dynamics of Improvised Interaction

Duration: March 2015 - February 2018

Coordinator: Ircam (Paris)

Other partners: Inria (Nancy), University of La Rochelle

Participants: Ken Déguernel, Nathan Libermann, Emmanuel Vincent

Abstract: The goal of this project was to design a music improvisation system able to listen to the other musicians, to improvise in their style, and to modify its improvisation according to their feedback in real time.

MULTISPEECH was responsible for designing a system able to improvise on multiple musical dimensions (melody, harmony) across multiple time scales.

9.2.2. ANR ArtSpeech

Project acronym: ArtSpeech

Project title: Synthèse articulatoire phonétique

Duration: October 2015 - March 2019

Coordinator: Yves Laprie

Other partners: Gipsa-Lab (Grenoble), IADI (Nancy), LPP (Paris)

Participants: Ioannis Douros, Yves Laprie, Anastasiia Tsukanova

Abstract: The objective is to synthesize speech from text via the numerical simulation of the human speech production processes, i.e. the articulatory, aerodynamic and acoustic aspects. Corpus based approaches have taken a hegemonic place in text to speech synthesis. They exploit very good acoustic quality speech databases while covering a high number of expressions and of phonetic contexts. This is sufficient to produce intelligible speech. However, these approaches face almost insurmountable obstacles as soon as parameters intimately related to the physical process of speech production have to be modified. On the contrary, an approach which rests on the simulation of the physical speech production process makes explicitly use of source parameters, anatomy and geometry of the vocal tract, and of a temporal supervision strategy. It thus offers direct control on the nature of the synthetic speech.

Static MRI acquisition of vowels (images plus acoustic signal) have been carried out this year and their exploitation started to explore the impact of the articulatory modeling and the plane wave assumption. Manual delineations of approximately 1000 images have been done and used to generate speech signals with articulatory copy synthesis.

9.2.3. ANR JCJC KAMoulox

Project acronym: KAMoulox

Project title: Kernel additive modelling for the unmixing of large audio archives

Duration: January 2016 - September 2019

Coordinator: Antoine Liutkus (Inria Zenith)

Participants: Mathieu Fontaine, Antoine Liutkus

Abstract: The objective is to develop the theoretical and applied tools required to embed audio denoising and separation tools in web-based audio archives. The applicative scenario is to deal with large audio archives, and more precisely with the notorious “Archives du CNRS — Musée de l’homme”, gathering about 50,000 recordings dating back to the early 1900s.

9.2.4. PIA2 ISITE LUE

Project acronym: ISITE LUE

Project title: Lorraine Université d’Excellence

Duration: starting in 2016

Coordinator: Univ. Lorraine

Participants: Ioannis Douros, Yves Laprie

Abstract: The initiative aims at developing and densifying the initial perimeter of excellence, within the scope of the social and economic challenges, so as to build an original model for a leading global engineering university, with a strong emphasis on technological research and education through research. For this, we have designed LUE as an “engine” for the development of excellence, by stimulating an original dialogue between knowledge fields.

MULTISPEECH is mainly concerned with challenge number 6: “Knowledge engineering”, i.e., engineering applied to the field of knowledge and language, which represent our immaterial wealth while being a critical factor for the consistency of future choices. This project funds the PhD thesis of Ioannis Douros.

9.2.5. E-FRAN METAL

Project acronym: E-FRAN METAL

Project title: Modèles Et Traces au service de l'Apprentissage des Langues

Duration: October 2016 - September 2020

Coordinator: Anne Boyer (LORIA)

Other partners: Interpsy, LISEC, ESPE de Lorraine, D@NTE (Univ. Versailles Saint Quentin), Sailendra SAS, ITOP Education, Rectorat.

Participants: Theo Biasutto-Lervat, Anne Bonneau, Vincent Colotte, Dominique Fohr, Denis Jouvet, Odile Mella, Slim Ouni, Anne-Laure Piat-Marchand, Elodie Gauthier, Thomas Girod

Abstract: METAL aims at improving the learning of languages (both written and oral components) through the development of new tools and the analysis of numeric traces associated with students' learning, in order to adapt to the needs and rythm of each learner.

MULTISPEECH is concerned by oral language learning aspects.

9.2.6. ANR VOCADOM

Project acronym: VOCADOM (http://vocadom.imag.fr/)

Project title: Robust voice command adapted to the user and to the context for ambient assisted living

Duration: January 2017 - December 2020

Coordinator: CNRS - LIG (Grenoble)

Other partners: Inria (Nancy), Univ. Lyon 2 - GREPS, THEORIS (Paris)

Participants: Dominique Fohr, Md Sahidullah, Sunit Sivasankaran, Emmanuel Vincent

Abstract: The goal of this project is to design a robust voice control system for smart home applications.

MULTISPEECH is responsible for wake-up word detection, overlapping speech separation, and speaker recognition.

9.2.7. ANR JCJC DiSCogs

Project acronym: DiSCogs

Project title: Distant speech communication with heterogeneous unconstrained microphone arrays

Duration: September 2018 – March 2022

Coordinator: Romain Serizel

Participants: Nicolas Furnon, Irina Illina, Romain Serizel, Emmanuel Vincent

Collaborators: Télécom ParisTech, 7sensing

Abstract: The objective is to solve fundamental sound processing issues in order to exploit the many devices equipped with microphones that populate our everyday life. The solution proposed is to apply machine learning methods based on deep learning to recast the problem of synchronizing devices at the signal level as a multi-view learning problem aiming at extracting complementary information from the devices at hand.

9.3. European Initiatives

9.3.1. FP7 & H2020 Projects

9.3.1.1. COMPRISE

Program: H2020 ICT-29-2018 (RIA)

Project acronym: COMPRISE

Project title: Cost-effective, Multilingual, Privacy-driven voice-enabled Services

Duration: Dec 2018- Nov 2021

Coordinator: Emmanuel Vincent

Other partners: Inria Magnet, Ascora GmbH, Netfective Technology SA, Rooter Analysis SL, Tilde SIA, University of Saarland

Participants: Irina Illina, Denis Jouvet, Emmanuel Vincent

Abstract: COMPRISE will define a fully private-by-design methodology and tools that will reduce the cost and increase the inclusiveness of voice interaction technologies.

9.3.2. Collaborations in European Programs, Except FP7 & H2020

9.3.2.1. AMIS

Program: CHIST-ERA

Project acronym: AMIS

Project title: Access Multilingual Information opinionS

Duration: Dec 2015- Nov 2018

Coordinator: Kamel Smaïli (LORIA)

Other partners: University of Avignon, University of Science and Technology Krakow, University of DEUSTO (Bilbao)

Participants: Dominique Fohr, Denis Jouvet, Odile Mella

Abstract: The idea of the project is to develop a multilingual help system of understanding without any human being intervention. This should help people understanding broadcasting news, presented in a foreign language and to compare it to a corresponding one available in the mother tongue of the user.

MULTISPEECH contributions concern mainly the speech recognition in French, English and Arabic videos.

9.4. International Initiatives

9.4.1. Inria International Partners

9.4.1.1. Informal International Partners

Jon Barker: University of Sheffield (UK)

Robust speech recognition [19]

Tomi Kinnunen: University of Eastern Finland (Finland)

Speaker verification and spoofing countermeasures for voice biometrics [47], [24], [32], [26], [41], [67]

Nicholas Evans: EURECOM (France)

Spoofing countermeasures for voice biometrics [47], [24], [32]

Hamid Eghbal-Zadeh: Johannes Kepler University (Austria)

Audio event detection [40]

Shinji Watanabe, Johns Hopkins University (USA)

Robust speech recognition [19]

Junichi Yamagishi, National Institute of Informatics (Japan)

Spoofing countermeasures for voice biometrics [47], [24], [32], [26]

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

Elected chair, Steering Committee of the Latent Variable Analysis and Signal Separation (LVA/ICA) conference series (E. Vincent)

General co-chair, 5th CHiME Speech Separation and Recognition Challenge, September 2018 (E. Vincent)

General co-chair, 5th International Workshop on Speech Processing in Everyday Environments, Hyderabad, India, September 2018 (E. Vincent)

Area chair, Latent Variable Analysis and Signal Separation (LVA/ICA), July 2018 (A. Deleforge)

Area Chair, INTERSPEECH, September 2018 (D. Jouvet)

Co-chair, special session "Geometry-Aware Auditory Scene Analysis" at ICASSP, April 2018 (A. Deleforge)

Co-chair, special session "Advances in Phase Retrieval and Application" at LVA/ICA, July 2018 (A. Deleforge)

Co-chair, special session "Distant speech processing for smart speakers" at EUSIPCO, September 2018 (R. Serizel)

Chair, IEEE signal processing cup (SPCup), 2019 (A. Deleforge)

10.1.1.2. Member of Organizing Committees

Steering Committee of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge series (E. Vincent)

Organizing Committe of Automatic Speaker Verification Spoofing and Countermeasures Challenge 2019 (ASVspoof 2019) (M. Sahidullah)

Organizing Committee of the DCASE challenge series, 2018 (R. Serizel)

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

Review chair, IEEE Technical Committee on Audio and Acoustic Signal Processing, responsible for organizing the review of the 443 papers submitted to the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) in the general AASP domain (E. Vincent)

10.1.2.2. Member of Conference Program Committees

Plate-Forme Intelligence Artificielle (PFIA), July 2018 (A. Bonneau, S. Ouni, R. Serizel)

International Conference on Multimedia and Network Information Systems (MISSI), September, 2018 (D. Jouvet)

International Conference on Speech and Computer (SPECOM), September 2018 (D. Jouvet)

International Conference on Text, Speech, and Dialogue (TSD), September 2018 (D. Jouvet)

10.1.2.3. Reviewer

ACLing 2018 - Arabic Computational Linguistics series (D. Jouvet)

APSIPA ASC 2018 - Asia-Pacific Signal and Information Processing Association - Annual Summit and Conference (M. Sahidullah)

CHiME 2018 - International Workshop on Speech Processing in Everyday Environments (E. Vincent)

EUSIPCO 2018 - European Signal Processing Conference (D. Jouvet, M. Sahidullah, R. Serizel)

ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing (A. Deleforge, I. Illina, D. Jouvet, S. Ouni, R. Serizel, E. Vincent)

ICPR 2018 - International Conference on Pattern Recognition (M. Sahidullah)

INTERSPEECH 2018 (A. Bonneau, I. Illina, S. Ouni, M. Sahidullah, E. Vincent)

IROS 2018 - International Conference on Intelligent Robots (A. Deleforge)

IWAENC 2018 - International Workshop on Acoustic Signal Enhancement (E. Vincent)

JEP 2018 - Journées d'Etudes sur la Parole (A. Bonneau, Y. Laprie, S. Ouni)

LVA/ICA 2018 - International Conference on Latent Variable Analysis and Signal Separation (E. Vincent)

MISSI 2018 - International Conference on Multimedia and Network Information Systems (D. Jouvet)

Speaker Odyssey 2018 - The Speaker and Language Recognition Workshop (M. Sahidullah)

SPECOM 2018 - International Conference on Speech and Computer (D. Jouvet)

SPEECH PROSODY 2018 (A. Bonneau, A. Deleforge, D. Jouvet)

SLT 2018 - IEEE Spoken Language Technology Workshop (D. Jouvet, M. Sahidullah, I. Illina)

10.1.3. Journal

10.1.3.1. Member of Editorial Boards

Speech Communication (D. Jouvet)

Speech Communication, special issue on Realism in Robust Speech and Language Processing (E. Vincent)

EURASIP Journal on Audio, Speech, and Music Processing (Y. Laprie)

10.1.3.2. Reviewer - Reviewing Activities

Computers in Biology and Medicine (M. Sahidullah)

Computer Speech & Language (M. Sahidullah)

Digital Signal Processing (M. Sahidullah)

Expert Systems With Applications (M. Sahidullah)

IEEE Access (M. Sahidullah)

IEEE/ACM Transactions on Audio, Speech, and Language Processing (A. Deleforge, M. Sahidullah, R. Serizel)

IEEE Journal of Selected Topics in Signal Processing (A. Deleforge, R. Serizel)

IEEE Signal Processing Letters (A. Deleforge, M. Sahidullah)

IEEE Transactions on Emerging Topics in Computational Intelligence (R. Serizel)

IEEE Transactions on Multimedia (S. Ouni)

Jasa Express Letters (Y. Laprie)

Journal of the Acoustical Society of America (Y. Laprie, S. Ouni)

Language Resources and Evaluation (S. Ouni)

Signal Processing Letters (D. Jouvet)

Speech Communication (A. Deleforge, M. Sahidullah)

10.1.4. Invited Talks

Deep learning for distant-microphone speech enhancement and recognition, Facebook AI Research, Paris, January 2018 (E. Vincent)

Multidimensional and multi-level learning of music structure for machine improvisation in the DYCI2 project, University Pompeu Fabra, Spain, February 2018 (E. Vincent)

Des difficultés aux troubles de l'apprentissage de la lecture – Impact sur l'estime de soi. Canopé, Nancy, February 2018 (A. Piquard-Kipffer)

A brief introduction to deep learning and its application to multichannel speech enhancement, Laboratoire des systèmes perceptifs, Paris, March 2018 (R. Serizel)

Speaker embeddings: from i-vector to x-vector and beyond in summer school on Machine Learning Applied to Speech Technology and Autonomous Agents, University of Eastern Finland, Finland, August 2018 (M. Sahidullah)

10.1.5. Leadership within the Scientific Community

Elected chair, ISCA Special Interest Group on Robust Speech Processing, until September 2018 (E. Vincent)

Secretary/Treasurer, executive member of AVISA (Auditory-VIsual Speech Association), an ISCA Special Interest Group (S. Ouni)

10.1.6. Scientific Expertise

Expertise of ANR project proposals (Y. Laprie)

Expertise of an ASTRID proposal (R. Serizel)

Expertise of a Netherlands Orgaadminisnisation for Scientific Research (NWO) proposal (R. Serizel)

Member of an expertise Committee for specific language disabilities (MDPH 54) (A. Piquard-Kipffer)

10.1.7. Research Administration

Head of the AM2I Scientific Pole of Université de Lorraine (Y. Laprie)

Vice Scientific Deputy of Inria Nancy - Grand Est (E. Vincent)

Member of Management board of Université de Lorraine (Y. Laprie)

Member of the Comité Espace Transfert of Inria Nancy - Grand Est (E. Vincent)

Member of the HCERES committee for the Laboratoire Psychologie de la Perception (E. Vincent)

Member of the HCERES committee for the LIMSI laboratory (D. Jouvet)

Vice-chair of the recruitement jury for an Associate Professor, Polytech Nancy (E. Vincent)

Member of the national recruitment jury for Inria Junior Research Scientists (E. Vincent)

Member of the Commission du personnel scientifique of Inria Nancy - Grand Est (R. Serizel)

Member of the Commission développement technologique of Inria Nancy - Grand Est (R. Serizel)

Member of "Commission paritaire" of Université de Lorraine (Y. Laprie)

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

DUT: I. Illina, Programming in Java, 150 hours, L1, University of Lorraine, France

DUT: I. Illina, Linux System, 65 hours, L1, University of Lorraine, France

DUT: I. Illina, Supervision of student projects and stages, 50 hours, L2, University of Lorraine, France

DUT: S. Ouni, Programming in Java, 24 hours, L1, University of Lorraine, France

DUT: S. Ouni, Web Programming, 24 hours, L1, University of Lorraine, France

DUT: S. Ouni, Graphical User Interface, 96 hours, L1, University of Lorraine, France

DUT: S. Ouni, Advanced Algorihms, 24 hours, L2, University of Lorraine, France

DUT: R. Serizel, Computer science basics, 90 hours, L1, University of Lorraine, France

DUT: R. Serizel, Introduction to office software applications, 18 hours, L2, University of Lorraine, France

DUT: R. Serizel, Multimedia and web applications, 20 hours, L1, University of Lorraine, France

DUT: R. Serizel, Multimedia content indexing, 20 hours, L2, University of Lorraine, France

Licence: A. Bonneau, Phonetics, 17 hours, L2, Ecole d'audioprothèse, University of Lorraine, France

Licence: A. Bonneau, Speech manipulations, 2 hours, L1, Ecole d'orthophonie, University of Lorraine, France

Licence: V. Colotte, C2i - Certificat Informatique et Internet, 50 hours, L1, University of Lorraine, France

Licence: V. Colotte, System, 45 hours, L3, University of Lorraine, France

Licence: O. Mella, Introduction to Web Programming, 30 hours, L1, University of Lorraine, France

Licence: O. Mella, Computer Networking, 98 hours, L2-L3, University of Lorraine, France

Licence: A. Piquard-Kipffer, Education Science, 32 hours, L1, Département d'Orthophonie, University of Lorraine, France

Licence: A. Piquard-Kipffer, Learning to Read, 34 hours, L2, Département d'Orthophonie, University of Lorraine, France

Licence: A. Piquard-Kipffer, Dyslexia, Dysorthographia, 12 hours, L3, Département d'Orthophonie, University of Lorraine, France

Licence: A. Piquard-Kipffer, Mathematics Didactics, 9 hours, L3, Département d'Orthophonie, University of Lorraine, France

Master: V. Colotte, Introduction to Speech Analysis and Recognition, 18 hours, M1, University of Lorraine, France

Master: V. Colotte, Integration project: multimodal interaction with Pepper, 10 hours, M2, University of Lorraine, France

Master: D. Jouvet, Speech recognition, 12 hours, M2 University of Lorraine, France

Master: D. Jouvet and S. Ouni, Multimodal oral comunication, 24 hours, M2, University of Lorraine, France

Master: Y. Laprie, Analysis, perception and automatic recognition of speech, 30 hours, M1, University of Lorraine, France

Master: O. Mella, Computer Networking, 76 hours, M1, University of Lorraine, France

Master: S. Ouni, Multimedia in Distributed Information Systems, 31 hours, M2, University of Lorraine, France

Master: E. Vincent and T. Biasutto-Lervat, Neural networks, 38 hours, M2, University of Lorraine

Master: A. Piquard-Kipffer, Dyslexia, Dysorthographia diagnosis, 6 hours, M1, Département d'Orthophonie, University of Lorraine, France

Master: A. Piquard-Kipffer, Deaf people and reading, 21 hours, Département d'Orthophonie, University of Lorraine, France

Master: A. Piquard-Kipffer, Psycholinguistics, 20 hours, Département d'Orthophonie, University Pierre et Marie Curie-Paris, France

Master: A. Piquard-Kipffer, French Language Didactics, 53 hours, ESPE, University of Lorraine, France

Master: A. Piquard-Kipffer, Psychology, 6 hours, M2, Department of Psychology, University of Lorraine, France

Continuous training: O. Mella, Computer science courses for secondary school teachers (ISN courses), 21 hours, ESPE, University of Lorraine, France

Continuous training: A. Piquard-Kipffer, Special Educational Needs, 53 hours, ESPE, University of Lorraine, France

Doctorat: A. Piquard-Kipffer, Language Pathology, 20 hours, EHESP, University of Sorbonne-Paris cité, France

Other: V. Colotte, Responsible for "Certificat Informatique et Internet" for the University of Lorraine, France (50000 students, 30 departments)

10.2.2. Supervision

PhD: Ken Déguernel, "Apprentissage de structures musicales en contexte d'improvisation", University of Lorraine, March 6, 2018, Emmanuel Vincent and Gérard Assayag (IRCAM) [10].

PhD: Baldwin Dumortier, "Contrôle acoustique d'un parc éolien", University of Lorraine, September 25, 2018, Emmanuel Vincent and Madalina Deaconu (Inria Tosca) [9].

PhD in progress: Amal Houidhek, "Synthèse paramétrique de parole arabe", December 2015, cotutelle, Vincent Colotte, Denis Jouvet and Zied Mnasri (ENIT, Tunisia).

PhD in progress: Imène Zangar, "Amélioration de la qualité de synthèse vocale par HMM pour la parole arabe", December 2015, codirection, Vincent Colotte, Denis Jouvet and Zied Mnasri (ENIT, Tunisia).

PhD in progress: Amine Menacer, "Traduction automatique de vidéos", May 2016, Kamel Smaïli and Denis Jouvet.

PhD in progress: Mathieu Fontaine, "Processus alpha-stable pour le traitement du signal", May 2016, Antoine Liutkus and Roland Badeau (Télécom ParisTech).

PhD in progress: Anastasiia Tsukanova, "Coarticulation modeling in articulatory synthesis", May 2016, Yves Laprie.

PhD in progress: Nathan Libermann, "Deep learning for musical structure analysis and generation", October 2016, Frédéric Bimbot (IRISA) and Emmanuel Vincent.

PhD in progress: Lauréline Perotin, "Séparation aveugle de sources sonores en milieu réverbérant", November 2016, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin (Orange).

PhD in progress: Théo Biasutto, "Multimodal coarticulation modeling: Towards the animation of an intelligible speaking head", December 2016, Slim Ouni.

PhD in progress: Sara Dahmani, "Modeling facial expressions to animate a realistic 3D virtual talking head", January 2017, Slim Ouni and Vincent Colotte.

PhD in progress: Guillaume Carbajal, "Apprentissage profond bout-en-bout pour le rehaussement de la parole", March 2017, Romain Serizel, Emmanuel Vincent, and Éric Humbert (Invoxia).

PhD in progress: Sunit Sivasankaran, "Exploiting contextual information in the speech processing chain", July 2017, Dominique Fohr and Emmanuel Vincent.

PhD in progress: Ioannis Douros, "Combining cineMRI and static MRI to analyze speech production", July 2017, Pierre-André Vuissoz (IADI) and Yves Laprie.

PhD in progress: Diego Di Carlo, "Estimating the Geometry of Audio Scenes Using Virtually-Supervised Learning", October 2017, Antoine Deleforge and Nancy Bertin.

PhD in progress: Lou Lee, "Du lexique au discours: les particules discursives en français", October 2017, Yvon Keromnes and Mathilde Dargnat (ATILF) and Denis Jouvet.

PhD in progress: Nicolas Turpault, "Deep learning for sound scene analysis in real environments", January 2018, Romain Serizel and Emmanuel Vincent.

PhD in progress: Raphaël Duroselle, “Adaptation de domaine par réseaux de neurones appliquée au traitement de la parole”, September 2018, Denis Jouvet and Irina Illina.

PhD in progress: Nicolas Furnon, “Deep-learning based speech enhancement with ad-hoc microphone arrays”, October 2018, Romain Serizel, Irina Illina and Slim Essid (Télécom ParisTech).

PhD in progress: Ajinkya Kulkarni, "Synthèse de parole expressive par apprentissage profond", October 2018, Vincent Colotte and Denis Jouvet.

PhD in progress: Manuel Pariente, “Deep learning-based phase-aware audio signal modeling and estimation”, October 2018, Antoine Deleforge and Emmanuel Vincent.

PhD in progress: Adrien Dufraux, “Leveraging noisy, incomplete, or implicit labels for automatic speech recognition”, November 2018, Emmanuel Vincent, Armelle Brun (LORIA), and Matthijs Douze (Facebook AI Research).

10.2.3. Juries

10.2.3.1. Participation in HDR and PhD juries

Participation in the Habilitation (HDR) Jury for Anthony Larcher (University of Le Mans, December 2018), D. Jouvet

Participation in the PhD jury of Anastasios Alexandridis (University of Crete, January 2018), E. Vincent.

Participation in the PhD jury of Marius Miron (University Pompeu Fabra, Spain, February 2018), E. Vincent, reviewer.

Participation in the PhD jury of Victor Bisot (Télécom ParisTech, March 2018), E. Vincent, reviewer.

Participation in the PhD jury of Boaz Schwarz (Bar-Ilan university, Israel, March 2018), A. Deleforge, reviewer.

Participation in the PhD jury of Jean-Rémy Gloaguen (École Centrale de Nantes, October 2018), E. Vincent, reviewer.

Participation in the PhD jury of Mathieu Andreux (Université Paris Sciences et Lettres, November 2018), E. Vincent, reviewer.

Participation in the PhD jury of Ming Xiu (Université de Strasbourg, December 2018), Y. Laprie.

10.2.3.2. Participation in other juries

Participation in CAFIPEMPF Jury - Master Learning Facilitator, Académie de Nancy-Metz & Université de Lorraine, April, May 2018, A. Piquard-Kipffer

Participation in the Competitive Entrance Examination into Speech-Language Pathology Departement, University of Lorraine, June 2018, A. Piquard-Kipffer.

10.3. Popularization

- Participation to the "Ada Lovelace Day" organized by Inria Nancy - Grand Est, October 2018 (A. Bonneau).

10.3.1. Articles and contents

- Interview for “GAFA : la voix est libre”, le Magazine de la Rédaction, *France Culture*, January 5, 2018 (E. Vincent)
- Interview for “Les enceintes connectées vont révolutionner vos vies et vous ne pourrez plus vous en passer”, *18h39.fr*, February 22, 2018 (E. Vincent)
- Interview for “Annecy 2018 : Dynalips, solution de lip-sync automatisée.”, *3DVF.com*, June 18, 2018 (S. Ouni)

10.3.2. Interventions

- Demonstration at Journée des métiers, Collège Péguy, le Chesnay, April 2018 (A. Piquard-Kipffer)
- Exhibition of an automatic lip-sync system at The Annecy International Animation Film Festival, Annecy, 12-15 June 2018 (S. Dahmani, V. Girard, T. Biasutto-Lervat, S. Ouni)
- Demonstration at Fête de la Science, University of Lorraine, October 12, 2018 (N. Turpault, R. Serizel, E. Vincent)
- Demonstration and participation in a panel discussion at Forum des Sciences Cognitives, Artem, Nancy, November 7, 2018 (N. Turpault, E. Vincent)
- Invitation to Serbia Science Festival 2018 (Belgrade) by the French institute of Serbia to give four popularization lectures and present a demonstration on "teaching robots to hear us", November 2018 (A. Deleforge).

11. Bibliography

Major publications by the team in recent years

[1] *Best Paper*
J. BARKER, E. VINCENT, N. MA, H. CHRISTENSEN, P. GREEN. *The PASCAL CHiME Speech Separation and Recognition Challenge*, in "Computer Speech and Language", February 2013, vol. 27, n° 3, pp. 621-633 [*DOI :* 10.1016/J.CSL.2012.10.004], https://hal.inria.fr/hal-00743529.

[2] A. BONNEAU, D. FOHR, I. ILLINA, D. JOUVET, O. MELLA, L. MESBAHI, L. OROSANU. *Gestion d'erreurs pour la fiabilisation des retours automatiques en apprentissage de la prosodie d'une langue seconde*, in "Traitement Automatique des Langues", 2013, vol. 53, n° 3, https://hal.inria.fr/hal-00834278

[3] D. JOUVET, D. FOHR. *Combining Forward-based and Backward-based Decoders for Improved Speech Recognition Performance*, in "InterSpeech - 14th Annual Conference of the International Speech Communication Association - 2013", Lyon, France, August 2013, https://hal.inria.fr/hal-00834282

[4] K. NATHWANI, J. A. MORALES-CORDOVILLA, S. SIVASANKARAN, I. ILLINA, E. VINCENT. *An extended experimental investigation of DNN uncertainty propagation for noise robust ASR*, in "5th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2017)", San Francisco, United States, March 2017, https://hal.inria.fr/hal-01446441

[5] A. A. NUGRAHA, A. LIUTKUS, E. VINCENT. *Multichannel audio source separation with deep neural networks*, in "IEEE/ACM Transactions on Audio, Speech, and Language Processing", June 2016, vol. 24, n° 10, pp. 1652-1664 [*DOI :* 10.1109/TASLP.2016.2580946], https://hal.inria.fr/hal-01163369

[6] A. OZEROV, E. VINCENT, F. BIMBOT. *A General Flexible Framework for the Handling of Prior Information in Audio Source Separation*, in "IEEE Transactions on Audio, Speech and Language Processing", May 2012, vol. 20, n° 4, pp. 1118 - 1133, 16, https://hal.archives-ouvertes.fr/hal-00626962

[7] A. PIQUARD-KIPFFER, C. BLONZ. *Je peux voir les mots que tu dis ! Histoire d'un projet*, in "13ème édition du Festival du film de chercheur CNRS 2012", Nancy, France, June 2012, https://hal.inria.fr/hal-01263907

[8] A. PIQUARD-KIPFFER, L. SPRENGER-CHAROLLES. *Predicting reading level at the end of Grade 2 from skills assessed in kindergarten: contribution of phonemic discrimination (Follow-up of 85 French-speaking children from 4 to 8 years old)*, in "Topics in Cognitive Psychology", 2013, https://hal.inria.fr/hal-00833951

Publications of the year

Doctoral Dissertations and Habilitation Theses

[9] B. DUMORTIER. *Acoustic control of wind farms*, Université de Lorraine, September 2018, https://tel.archives-ouvertes.fr/tel-01897853

[10] K. DÉGUERNEL. *Learning of musical structures in the context of improvisation*, Université de Lorraine, March 2018, https://tel.archives-ouvertes.fr/tel-01735308

Articles in International Peer-Reviewed Journals

[11] N. BERTIN, E. CAMBERLEIN, R. LEBARBENCHON, E. VINCENT, S. SIVASANKARAN, I. ILLINA, F. BIMBOT. *VoiceHome-2, an extended corpus for multichannel speech processing in real homes*, in "Speech Communication", 2018, https://hal.inria.fr/hal-01923108

[12] K. DÉGUERNEL, E. VINCENT, G. ASSAYAG. *Probabilistic Factor Oracles for Multidimensional Machine Improvisation*, in "Computer Music Journal", June 2018, vol. 42, n° 2, pp. 52-66 [*DOI :* 10.1162/COMJ_A_00460], https://hal.inria.fr/hal-01693750

[13] A. HOUIDHEK, V. COLOTTE, Z. MNASRI, D. JOUVET. *Evaluation of speech unit modelling for HMM-based speech synthesis for Arabic*, in "International Journal of Speech Technology", November 2018, pp. 1-12 [*DOI :* 10.1007/s10772-018-09558-6], https://hal.inria.fr/hal-01936963

[14] D. JOUVET, D. LANGLOIS, M. A. MENACER, D. FOHR, O. MELLA, K. SMAÏLI. *Adaptation of speech recognition vocabularies for improved transcription of YouTube videos*, in "Journal of International Science and General Applications", March 2018, vol. 1, n° 1, pp. 1-9, https://hal.archives-ouvertes.fr/hal-01873801

[15] K. NATHWANI, E. VINCENT, I. ILLINA. *DNN Uncertainty Propagation using GMM-Derived Uncertainty Features for Noise Robust ASR*, in "IEEE Signal Processing Letters", January 2018 [*DOI :* 10.1109/LSP.2018.2791534], https://hal.inria.fr/hal-01680658

[16] S. OUNI, G. GRIS. *Dynamic Lip Animation from a Limited number of Control Points: Towards an Effective Audiovisual Spoken Communication*, in "Speech Communication", February 2018, vol. 96, pp. 49-57 [*DOI :* 10.1016/J.SPECOM.2017.11.006], https://hal.inria.fr/hal-01631397

[17] Z. WANG, E. VINCENT, R. SERIZEL, Y. YAN. *Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments*, in "Computer Speech and Language", May 2018, vol. 49, pp. 37-51 [*DOI :* 10.1016/J.CSL.2017.11.003], https://hal.inria.fr/hal-01634449

International Conferences with Proceedings

[18] B. ABDULLAH, I. ILLINA, D. FOHR. *Dynamic Extension of ASR Lexicon Using Wikipedia Data*, in "IEEE Workshop on Spoken and Language Technology (SLT)", Athènes, Greece, Proceedings of IEEE SLT, December 2018, https://hal.archives-ouvertes.fr/hal-01874495

[19] J. BARKER, S. WATANABE, E. VINCENT, J. TRMAL. *The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines*, in "Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association", Hyderabad, India, September 2018, https://arxiv.org/abs/1803.10609 , https://hal.inria.fr/hal-01744021

[20] K. BARTKOVA, D. JOUVET. *Analysis of prosodic correlates of emotional speech data*, in "ExLing 2018 - 9th Tutorial and Research Workshop on Experimental Linguistics", Paris, France, August 2018, https://hal.inria.fr/hal-01889932

[21] T. BIASUTTO– LERVAT, S. OUNI. *Phoneme-to-Articulatory mapping using bidirectional gated RNN*, in "Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association", Hyderabad, India, September 2018, https://hal.inria.fr/hal-01862587

[22] A. BONNEAU. *Impact of fluency and segmental categorization in L2: the case of French final fricatives uttered by German speakers*, in "Speech Prosody 2018", Poznan, Poland, June 2018 [*DOI :* 10.21437/SPEECHPROSODY.2018-189], https://hal.inria.fr/hal-01926657

[23] G. CARBAJAL, R. SERIZEL, E. VINCENT, E. HUMBERT. *Multiple-input neural network-based residual echo suppression*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Canada, April 2018, pp. 1-5, https://hal.inria.fr/hal-01723630

[24] H. DELGADO, M. TODISCO, M. SAHIDULLAH, N. EVANS, T. KINNUNEN, K. A. LEE, J. YAMAGISHI. *ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements*, in "Odyssey 2018 - The Speaker and Language Recognition Workshop", Les Sables d'Olonne, France, June 2018, https://hal.inria.fr/hal-01880206

[25] D. DI CARLO, A. LIUTKUS, K. DÉGUERNEL. *Interference reduction on full-length live recordings*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech, and Signal Processing", Calgary, Canada, IEEE, April 2018, pp. 736-740 [*DOI :* 10.1109/ICASSP.2018.8462621], https://hal.inria.fr/hal-01713889

[26] F. FANG, J. YAMAGISHI, I. ECHIZEN, M. SAHIDULLAH, T. KINNUNEN. *Transforming acoustic characteristics to deceive playback spoofing countermeasures of speaker verification systems*, in "WIFS 2018 - IEEE International Workshop on Information Forensics and Security", Hong Kong, Hong Kong SAR China, December 2018, https://hal.inria.fr/hal-01889910

[27] M. FONTAINE, F.-R. STÖTER, A. LIUTKUS, U. SIMSEKLI, R. SERIZEL, R. BADEAU. *Multichannel Audio Modeling with Elliptically Stable Tensor Decomposition*, in "LVA ICA 2018 - 14th International Conference on Latent Variable Analysis and Signal Separation", Surrey, United Kingdom, July 2018, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01766795

[28] B. GARCIA-ZAPIRAIN, C. CASTILLO, A. BADIOLA, S. ZAHIA, A. MENDEZ, D. LANGLOIS, D. JOUVET, J.-M. TORRES-MORENO, M. LESZCZUK, K. SMAÏLI. *A Proposed Methodology for Subjective Evaluation of Video and Text Summarization*, in "MISSI 2018 - 11th edition of the International Conference on Multimedia and Network Information Systems", Wrocław, Poland, Advances in Intelligent Systems and Computing, Springer, September 2018, vol. 833, pp. 396-404 [*DOI :* 10.1007/978-3-319-98678-4_40], https://hal.archives-ouvertes.fr/hal-01873685

[29] M. L. GREGA, K. SMAÏLI, M. LESZCZUK, C.-E. GONZÁLEZ-GALLARDO, J.-M. TORRES-MORENO, E. LINHARES PONTES, D. FOHR, O. MELLA, M. A. MENACER, D. JOUVET. *An Integrated AMIS Prototype for Automated Summarization and Translation of Newscasts and Reports*, in "MISSI 2018 - 11th International Conference on Multimedia and Network Information Systems", Wroclaw, Poland, K. CHOROŚ, M. KOPEL, E. KUKLA, A. SIEMIŃSKI (editors), Springer, September 2018, vol. 833, pp. 415-423 [*DOI :* 10.1007/978-3-319-98678-4_42], https://hal.archives-ouvertes.fr/hal-01873680

[30] A. HOUIDHEK, V. COLOTTE, Z. MNASRI, D. JOUVET. *DNN-Based Speech Synthesis for Arabic: Modelling and Evaluation*, in "SLSP 2018 - 6th International Conference on Statistical Language and Speech Processing", Mons, Belgium, October 2018, https://hal.inria.fr/hal-01904512

[31] N. KERIVEN, A. DELEFORGE, A. LIUTKUS. *Blind Source Separation Using Mixtures of Alpha-Stable Distributions*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Canada, IEEE, April 2018, pp. 771-775, https://arxiv.org/abs/1711.04460 [*DOI :* 10.1109/ICASSP.2018.8462095], https://hal.inria.fr/hal-01633215

[32] T. KINNUNEN, K. A. LEE, H. DELGADO, N. EVANS, M. TODISCO, M. SAHIDULLAH, J. YAMAGISHI, D. A. REYNOLDS. *t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification*, in "Speaker Odyssey 2018 The Speaker and Language Recognition Workshop", Les Sables d'Olonne, France, June 2018, https://hal.inria.fr/hal-01880306

[33] Y. LAPRIE, B. ELIE, A. TSUKANOVA, P.-A. VUISSOZ. *Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech*, in "EUSIPCO 2018 - 26th European Signal Processing Conference", Rome, Italy, September 2018, https://hal.inria.fr/hal-01921928

[34] L. LEE, K. BARTKOVA, M. DARGNAT, D. JOUVET. *Prosodic and Pragmatic Values of Discourse Particles in French*, in "ExLing 2018 - 9th Tutorial and Research Workshop on Experimental Linguistics", Paris, France, August 2018, https://hal.inria.fr/hal-01889925

[35] A. LIUTKUS, C. ROHLFING, A. DELEFORGE. *Audio source separation with magnitude priors: the BEADS model*, in "ICASSP: International Conference on Acoustics, Speech and Signal Processing", Calgary, Canada, Signal Processing and Artificial Intelligence: Changing the World, April 2018, pp. 1-5 [*DOI :* 10.1109/ICASSP.2018.8462515], https://hal.inria.fr/hal-01713886

[36] H. PEIC TUKULJAC, A. DELEFORGE, R. GRIBONVAL. *MULAN: A Blind and Off-Grid Method for Multichannel Echo Retrieval*, in "NeurIPS 2018 - Thirty-second Conference on Neural Information Processing Systems", Montréal, Canada, December 2018, pp. 1-11, https://arxiv.org/abs/1810.13338 , https://hal.inria.fr/hal-01906385

[37] L. PEROTIN, R. SERIZEL, E. VINCENT, A. GUÉRIN. *CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector*, in "IWAENC 2018 - 16th International Workshop on Acoustic Signal Enhancement", Tokyo, Japan, September 2018, https://hal.inria.fr/hal-01840453

[38] L. PEROTIN, R. SERIZEL, E. VINCENT, A. GUÉRIN. *Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings*, in "43rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2018)", Calgary, Canada, April 2018, https://hal.inria.fr/hal-01699759

[39] R. SCHEIBLER, D. DI CARLO, A. DELEFORGE, I. DOKMANIĆ. *Separake: Source Separation with a Little Help From Echoes*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Canada, April 2018, https://hal.inria.fr/hal-01909531

[40] R. SERIZEL, N. TURPAULT, H. EGHBAL-ZADEH, A. PARAG SHAH. *Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments*, in "Workshop on Detection and Classification of Acoustic Scenes and Events", Woking, United Kingdom, November 2018, https://arxiv.org/abs/1807.10501 - Submitted to DCASE2018 Workshop, https://hal.inria.fr/hal-01850270

[41] S. SIERANOJA, M. SAHIDULLAH, T. KINNUNEN, J. KOMULAINEN, A. HADID. *Audiovisual Synchrony Detection with Optimized Audio Features*, in "ICSIP 2018 - 3rd International Conference on Signal and Image Processing", Shenzhen, China, July 2018, https://hal.inria.fr/hal-01889918

[42] S. SIVASANKARAN, B. M. L. SRIVASTAVA, S. SITARAM, K. BALI, M. CHOUDHURY. *Phone Merging for Code-switched Speech Recognition*, in "Third Workshop on Computational Approaches to Linguistic Code-switching", Melbourne, Australia, collocated with ACL 2018 , July 2018, https://hal.inria.fr/hal-01800466

[43] S. SIVASANKARAN, E. VINCENT, D. FOHR. *Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment*, in "Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association", Hyderabad, India, September 2018, https://hal.archives-ouvertes.fr/hal-01817519

[44] *Best Paper*
K. SMAÏLI, D. FOHR, C. GONZÁLEZ-GALLARDO, M. GREGA, L. JANOWSKI, D. JOUVET, A. KOMOROWSKI, A. KOZBIAL, D. LANGLOIS, M. LESZCZUK, O. MELLA, M. A. MENACER, A. MENDEZ, E. LINHARES PONTES, E. SANJUAN, D. SWIST, J.-M. TORRES-MORENO, B. GARCIA-ZAPIRAIN. *A First Summarization System of a Video in a Target Language*, in "MISSI 2018 - 11th edition of the International Conference on Multimedia and Network Information Systems", Wrocław, Poland, September 2018, pp. 1-12, https://hal.archives-ouvertes.fr/hal-01819720.

[45] M. STRAUSS, P. MORDEL, V. MIGUET, A. DELEFORGE. *DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization*, in "IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018)", Madrid, Spain, October 2018, https://hal.inria.fr/hal-01854878

[46] L. TERISSI, G. SAD, M. CERDA, S. OUNI, R. GALVEZ, J. B. GÓMEZ, B. GIRAU, N. HITSCHFELD-KAHLER. *A French-Spanish Multimodal Speech Communication Corpus Incorporating Acoustic Data, Facial, Hands and Arms Gestures Information*, in "Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association", Hyderabad, India, September 2018, https://hal.inria.fr/hal-01862585

[47] M. TODISCO, H. DELGADO, K. A. LEE, M. SAHIDULLAH, N. EVANS, T. KINNUNEN, J. YAMAGISHI. *Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion*, in "Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association", Hyderabad, India, ISCA, September 2018 [*DOI :* 10.21437/INTERSPEECH.2018-2289], https://hal.inria.fr/hal-01889934

[48] Z. WANG, J. LI, Y. YAN, E. VINCENT. *Semi-supervised learning with deep neural networks for relative transfer function inverse regression*, in "ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Canada, April 2018, https://hal.inria.fr/hal-01797886

[49] I. ZANGAR, Z. MNASRI, V. COLOTTE, D. JOUVET, A. HOUIDHEK. *Duration modeling using DNN for Arabic speech synthesis*, in "Speech Prosody 2018 - Proceedings of the 9th International Conference on Speech Prosody are now available!", Poznań, Poland, June 2018, https://hal.inria.fr/hal-01889917

National Conferences with Proceedings

[50] N. LIBERMANN, F. BIMBOT, E. VINCENT. *Exploration de dépendances structurelles mélodiques par réseaux de neurones récurrents*, in "JIM 2018 - Journées d'Informatique Musicale", Amiens, France, May 2018, pp. 81-86, https://hal.archives-ouvertes.fr/hal-01791381

Conferences without Proceedings

[51] M. VACHER, E. VINCENT, M.-E. BOBILLIER CHAUMON, T. JOUBERT, F. PORTET, D. FOHR, S. CAFFIAU, T. DESOT. *The VocADom Project: Speech Interaction for Well-being and Reliance Improvement*, in "MobileHCI 2018 - 20th International Conference on Human-Computer Interaction with Mobile Devices and Services", Barcelona, Spain, September 2018, https://hal.archives-ouvertes.fr/hal-01830217

Scientific Books (or Scientific Book chapters)

[52] A. DELEFORGE, A. SCHMIDT, W. KELLERMANN. *Audio-Motor Integration for Robot Audition*, in "Multimodal Behavior Analysis in the Wild", Academic Press, November 2018, pp. 1-27, https://hal.inria.fr/hal-01929388

[53] C. FÉVOTTE, E. VINCENT, A. OZEROV. *Single-channel audio source separation with NMF: divergences, constraints and algorithms*, in "Audio Source Separation", Springer, March 2018, https://hal.inria.fr/hal-01631185

[54] T. GERKMANN, E. VINCENT. *Spectral masking and filtering*, in "Audio source separation and speech enhancement", E. VINCENT, T. VIRTANEN, S. GANNOT (editors), Wiley, August 2018, https://hal.inria.fr/hal-01881425

[55] A. A. NUGRAHA, A. LIUTKUS, E. VINCENT. *Deep neural network based multichannel audio source separation*, in "Audio Source Separation", Springer, March 2018, https://hal.inria.fr/hal-01633858

[56] A. OZEROV, C. FÉVOTTE, E. VINCENT. *An introduction to multichannel NMF for audio source separation*, in "Audio Source Separation", Signals and Communication Technology, Springer, March 2018, https://hal.inria.fr/hal-01631187

[57] M. SAHIDULLAH, H. DELGADO, M. TODISCO, T. KINNUNEN, N. EVANS, J. YAMAGISHI, K.-A. LEE. *Introduction to Voice Presentation Attack Detection and Recent Advances*, in "Handbook of Biometric Anti-Spoofing: Presentation Attack Detection", S. MARCEL, M. S. NIXO, J. FIERREZ, N. EVANS (editors), Advances in Computer Vision and Pattern Recognition, Springer, 2019, pp. 321-361, https://hal.inria.fr/hal-01974528

[58] A. TSUKANOVA, B. ELIE, Y. LAPRIE. *Articulatory Speech Synthesis from Static Context-Aware Articulatory Targets*, in "Studies on Speech Production", Q. FANG, J. DANG, P. PERRIER, J. WEI, L. WANG, N. YAN (editors), Lecture Notes in Computer Science, Springer, 2018, n° 10733, pp. 37-47, Revised Selected Papers of the 11th International Seminar, ISSP 2017, Tianjin, China, October 16-19, 2017 [*DOI :* 10.1007/978-3-030-00126-1_4], https://hal.archives-ouvertes.fr/hal-01937950

[59] E. VINCENT, S. GANNOT, T. VIRTANEN. *Acoustics - Spatial properties*, in "Audio source separation and speech enhancement", E. VINCENT, T. VIRTANEN, S. GANNOT (editors), Wiley, August 2018, https://hal.inria.fr/hal-01881423

[60] E. VINCENT, S. GANNOT, T. VIRTANEN. *Introduction*, in "Audio source separation and speech enhancement", E. VINCENT, T. VIRTANEN, S. GANNOT (editors), Wiley, August 2018, https://hal.inria.fr/hal-01881422

[61] E. VINCENT, T. VIRTANEN, S. GANNOT. *Perspectives*, in "Audio source separation and speech enhancement", E. VINCENT, T. VIRTANEN, S. GANNOT (editors), Wiley, August 2018, https://hal.inria.fr/hal-01881424

[62] T. VIRTANEN, E. VINCENT, S. GANNOT. *Time-frequency processing - Spectral properties*, in "Audio source separation and speech enhancement", E. VINCENT, T. VIRTANEN, S. GANNOT (editors), Wiley, August 2018, https://hal.inria.fr/hal-01881426

[63] M. ZITT, A. LELU, M. CADOT, G. CABANAC. *Bibliometric delineation of scientific fields*, in "Handbook of Science and Technology Indicators", W. GLÄNZEL, H. F. MOED, U. SCHMOCH, M. THELWALL (editors), Handbook of Science and Technology Indicators, Springer International Publishing, 2018 [*DOI :* 10.1007/978-3-030-02511-3], https://hal.archives-ouvertes.fr/hal-01942528

Books or Proceedings Editing

[64] E. VINCENT, T. VIRTANEN, S. GANNOT (editors). *Audio source separation and speech enhancement*, Wiley, August 2018, 504 p. [*DOI :* 10.1002/9781119279860], https://hal.inria.fr/hal-01881431

Research Reports

[65] M. CADOT, A. LELU, M. ZITT. *Benchmarking seventeen clustering methods on a text dataset: Comparaison empirique de dix-sept méthodes de classification non-supervisée sur un corpus textuel*, LORIA, March 2018, Version française en fichier complémentaire, https://hal.archives-ouvertes.fr/hal-01532894

Patents and standards

[66] S. OUNI, G. GRIS. *Image processing device*, March 2018, n° US2018/0061109 A1, https://hal.inria.fr/hal-01862639

Other Publications

[67] T. KINNUNEN, R. G. HAUTAMÄKI, V. VESTMAN, M. SAHIDULLAH. *Can We Use Speaker Recognition Technology to Attack Itself? Enhancing Mimicry Attacks Using Automatic Target Speaker Selection*, 2018, (A slightly shorter version) has been submitted to IEEE ICASSP 2019, https://hal.inria.fr/hal-01937767

[68] L. PEROTIN, R. SERIZEL, E. VINCENT, A. GUÉRIN. *CRNN-based multiple DoA estimation using Ambisonics acoustic intensity features*, July 2018, Submited to the IEEE Journal of Selected Topics in Signal Processing, Special Issue on Acoustic Source Localization and Tracking in Dynamic Real-life Scenes, https://hal.inria.fr/hal-01839883

References in notes

[69] L. SPRENGER-CHAROLLES, P. COLÉ, D. BÉCHENNEC, A. KIPFFER-PIQUARD. *French normative data on reading and related skills from EVALEC, a new computerized battery of tests (end Grade 1, Grade 2, Grade 3, and Grade 4)*, in "European Review of Applied Psychology / Revue Européenne de Psychologie Appliquée", 2005, n° 55, pp. 157-186, https://hal.inria.fr/inria-00184979